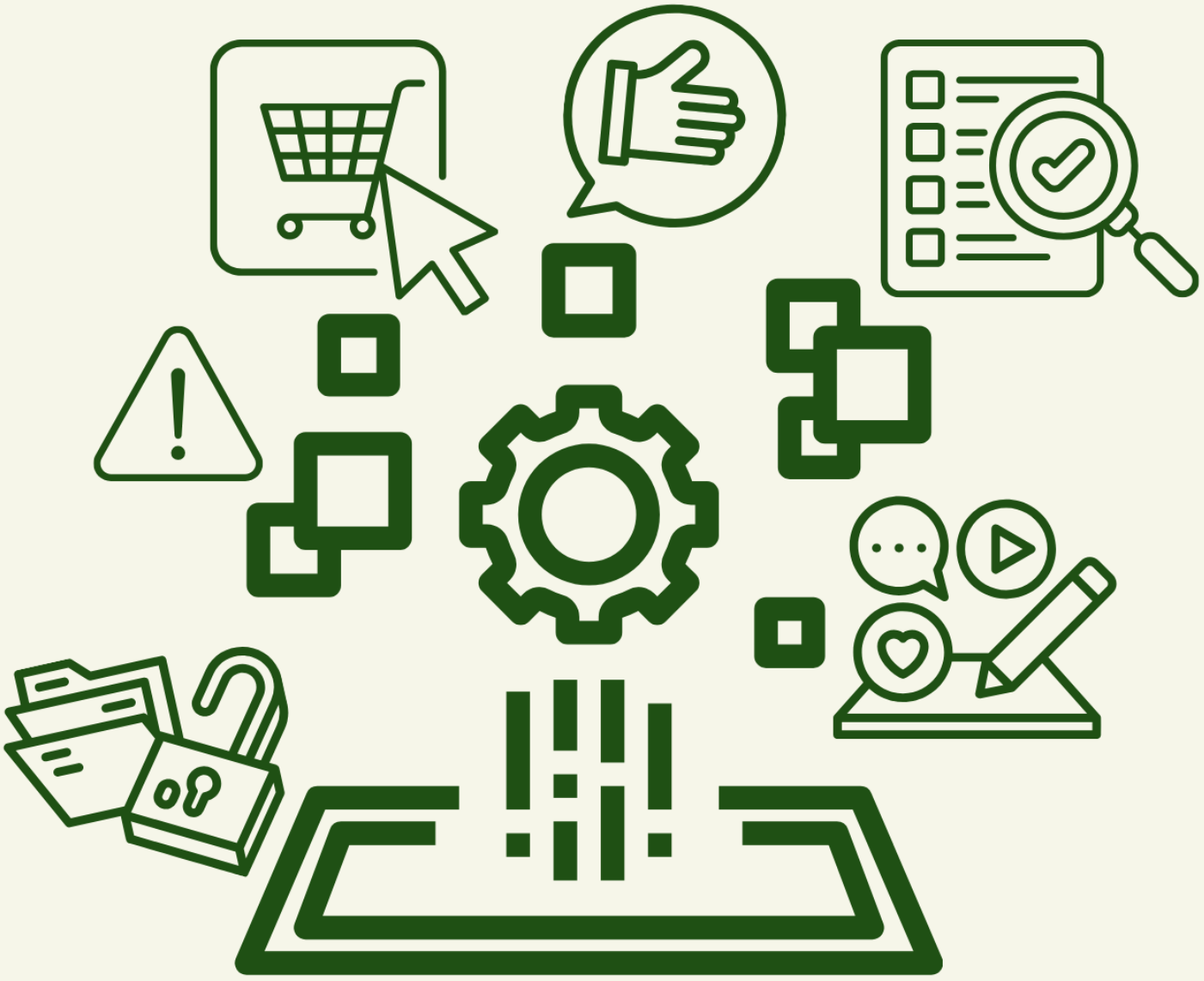
Platform Transparency under the EU's Digital Services Act



Opportunities and Challenges for the Global South

Published by
Centre for Communication Governance
National Law University Delhi
Sector 14, Dwarka, New Delhi – 110078

ISBN: 978-93-84272-56-2

Patrons: Professor (Dr.) G.S. Bajpai (Vice Chancellor, NLUD), Professor (Dr.) Ruhi Paul (Registrar, NLUD)

Faculty Director, CCG: Dr. Daniel Mathew

Executive Director, CCG: Jhalak M. Kakkar

Supported by: Global Network Initiative

Authors: Tavishi, Shobhit S.

Editing and Supervision: Shashank Mohan

Review: Shashank Mohan, Jhalak M. Kakkar

Conceptualisation: Tavishi, Shobhit S., Shashank Mohan, Jhalak M. Kakkar

Research Assistant: Ishita Tulsyan

Design: Gopika P.

[all names are in reverse alphabetical order]

Suggested Citation: Tavishi and Shobhit S., 'Platform Transparency under the EU's Digital Services Act: Opportunities and Challenges for the Global South' (Centre for Communication Governance, National Law University Delhi 2025)

PLATFORM TRANSPARENCY UNDER THE EU'S DIGITAL SERVICES ACT: Opportunities And Challenges for the Global South

Centre for Communication Governance

National Law University Delhi

FOREWORD

The digital public sphere has rapidly become one of the most consequential arenas in contemporary society. Digital platforms that were once celebrated for enabling unprecedented connectivity and the free flow of ideas are now at the centre of critical debates about accountability, fairness, and governance. Their reach and influence have reshaped political participation, social discourse, and economic opportunity across the world, while also generating significant risks-from the spread of misinformation and hate speech to threats against individual privacy and the integrity of democratic processes. These developments underscore a critical paradox: digital platforms have deepened opportunities for engagement and innovation, but in the absence of effective governance, they have also amplified harms that disproportionately affect vulnerable communities, especially in the Global South.

It is in this context that the European Union's Digital Services Act (DSA) represents a landmark development. The DSA seeks to establish transparency as a central principle of platform regulation, requiring disclosures on recommender systems, online advertising, risk assessments, audits, researcher access to data, and content moderation. This legislation embodies an ambitious attempt to rebalance the power dynamics between platforms, regulators, and the public by reducing information asymmetries and institutionalising accountability mechanisms. While the DSA is rooted in Europe's legal and political traditions, its scope and ambition ensure that it will reverberate globally, shaping regulatory debates well beyond EU's borders.

For scholars, policymakers, and civil society in the Global South, the DSA provides both inspiration and caution. On the one hand, it offers a comprehensive framework that articulates what meaningful transparency could look like, providing concrete standards against which platforms can be held accountable. On the other, it highlights the significant institutional and political capacity required to make such transparency meaningful. Many countries in Asia, Africa, and Latin America face acute challenges in this regard: regulators often lack resources, civil society is constrained in its ability to scrutinise platform behaviour, and political environments may inadvertently curtail legitimate speech.

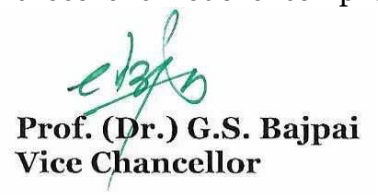
The report Platform Transparency under the EU's Digital Services Act: Opportunities and Challenges for the Global South, produced by the Centre for Communication Governance at the National Law University Delhi, makes an important contribution to this conversation. Drawing on detailed analysis of the DSA's provisions, it explores how transparency mechanisms might function if adapted to the specific contexts of the Global South. It carefully evaluates both the opportunities-such as enabling more informed policy interventions, empowering researchers, and improving public accountability-and the risks, including regulatory capture, institutional overreach, and the reinforcement of existing inequalities in platform governance. In doing so, the report provides a nuanced framework for understanding the complexities of transnational regulatory convergence in the digital age.

This work is particularly timely for several reasons. First, the Global South is not peripheral to the digital ecosystem; it represents the fastest-growing base of internet users worldwide. The challenges of content moderation in low-resource languages, the spread of disinformation in fragile democracies, and the labour conditions of outsourced moderation workers are not abstract issues but pressing realities for millions. Any global framework for platform governance that does not adequately account for these contexts risks reinforcing structural inequities. Second, as governments in the Global South increasingly engage with digital regulation-whether through data protection laws, intermediary liability frameworks, or online safety legislation-there is an urgent need for comparative analysis that draws lessons from other jurisdictions while remaining sensitive to local political and institutional conditions.

The broader lesson that emerges from this report is that transparency, while necessary, is not sufficient. Disclosures, audits, and risk assessments acquire meaning only when they are embedded within institutional ecosystems that can interpret, challenge, and act upon them. In the Global South, where such ecosystems are often fragile, the path towards meaningful transparency will require investments in institutional capacity, strengthening of civil society, and safeguarding of academic freedom.

As Vice Chancellor of the National Law University Delhi, I commend the Centre for Communication Governance's leadership in producing scholarship that engages directly with these pressing global issues. This report exemplifies the Centre's commitment to rigorous, policy-relevant research that bridges the worlds of academia, governance, and civil society. By situating the DSA within the broader context of platform regulation in the Global South, the report has advanced a discourse that is not only comparative but also normatively grounded in the principles of accountability, inclusivity, and respect for fundamental rights and values.

I envision this report will serve as a resource for a wide range of stakeholders. For policymakers, it offers concrete insights into the design and implementation of transparency frameworks that balance accountability with the protection of rights. For civil society organisations, it provides analytical tools to advocate for greater platform responsibility in ways that are contextually informed. For researchers and students, it opens new avenues of inquiry into the complex intersections between technology, law, and society. And for platforms themselves, it presents an opportunity to reflect on their global responsibilities and the need to move beyond a Eurocentric model of compliance.



Prof. (Dr.) G.S. Bajpai
Vice Chancellor

ACKNOWLEDGEMENTS

The successful culmination of this report owes much to the steadfast support provided by the Global Network Initiative (GNI). We sincerely appreciate their invaluable contributions in making this research possible. In particular, we are grateful to Jason Pielemeier, Elonnai Hickok, Hilary Ross, and Ramsha Jahangir from GNI for their consistent engagement and support, which has been instrumental in the publication of this report.

We express our gratitude to the National Law University of Delhi (NLUD) for its ongoing institutional support. We are especially thankful for the leadership and guidance of the Vice Chancellor, Prof. (Dr.) G. S. Bajpai, and the efforts of the Registrar, Prof. (Dr.) Ruhi Paul, in bringing this research to fruition. Our sincere thanks also go to Dr. Daniel Mathew, Faculty Director at CCG, for his thoughtful direction and steady counsel.

We are immensely thankful to Jhalak M. Kakkar, Executive Director at CCG, for her continued encouragement, mentorship and support in the development of this report. We are also deeply grateful to Shashank Mohan, Associate Director, for shaping this report with his wholehearted guidance and meticulous feedback at every stage. Special appreciation is due to the ever-reliable and ever-patient Suman Negi, Preeti Bhandari and Mahesh Singh for their tireless contributions to all the work we do at CCG. We are also grateful to our present and former colleagues at CCG, whose collaboration and enthusiasm has been a constant source of motivation.

We extend our appreciation to Vincent Hoffman, Nicolás Zara, and Helani Galpaya for generously offering their time and expertise, which greatly enhanced the quality and analytical depth of this report.

We are also extremely grateful to Nicolo Zingales, Prof. Kyung Sin Park, Karthik Nachiappan, and Francisco Brito Cruz for their thoughtful insights, which enriched our research. Finally, we express our gratitude to Prof. (Dr.) Joris van Hoboken for engaging patiently with us during our presentation of an earlier draft, and offering suggestions that encouraged us to expand our lines of inquiry.

ABOUT THE NATIONAL LAW UNIVERSITY, DELHI

The National Law University Delhi is one of the leading law universities of India based in the capital city of India. Established in 2008 (by Act. No. 1 of 2009), the University is ranked second in the National Institutional Ranking Framework for the last five years. Dynamic in vision and robust in commitment, the University has shown terrific promise to become a world-class institution in a very short span of time. It follows a mandate to transform and redefine the process of legal education. The primary mission of the University is to create lawyers who will be professionally competent, technically sound and socially relevant, and will not only enter the Bar and the Bench but also be equipped to address the imperatives of the new millennium and uphold constitutional values.

The University aims to evolve and impart comprehensive and interdisciplinary legal education which will promote legal and ethical values, while fostering the rule of law. The University offers a five-year integrated B.A., LL. B (Hons.) and one-year postgraduate masters in law (LL.M), along with professional programs, diploma and certificate courses for both lawyers and non-lawyers. The University has made tremendous contributions to public discourse on law through pedagogy and research.

Over the last decade, the University has established many specialised research centres, and this includes the Centre for Communication Governance, the Centre for Innovation, Intellectual Property and Competition, the Centre for Corporate Law and Governance, the Centre for Criminology and Victimology, and Project 39A. The University has made submissions, recommendations, and worked in advisory/consultant capacities with government entities, universities in India and abroad, think tanks, private sector organisations, and international organisations. The University works in collaboration with other international universities on various projects and has established MoU's with several other academic institutions.

ABOUT THE CENTRE FOR COMMUNICATION GOVERNANCE

The Centre for Communication Governance at the National Law University Delhi (CCG) was established in 2013 to ensure that Indian legal education establishments engage more meaningfully with information technology law and policy and contribute to improved governance and policy making. CCG is the only academic research centre dedicated to undertaking rigorous academic research on information technology law and policy in India. It has, in a short span of time, become a leading institution in Asia. Through its academic and policy research, CCG engages meaningfully with policy-making in India by participating in public consultations, contributing to parliamentary committees and other consultation groups, and holding seminars, courses and workshops for capacity building of different stakeholders in the technology law and policy domain. CCG works across issues such as privacy and data governance, platform governance, and emerging technologies.

CCG has built an extensive network and works with a range of international academic institutions and policy organisations. These include the United Nations Development Programme, Law Commission of India, NITI Aayog, various Indian government ministries and regulators, International Telecommunications Union, UNGA WSIS, Paris Call, Berkman Klein Center for Internet and Society at Harvard University, the Center for Internet and Society at Stanford University, Columbia University's Global Freedom of Expression and Information Jurisprudence Project, the Hans Bredow Institute at the University of Hamburg, the Programme in Comparative Media Law and Policy at the University of Oxford, the Annenberg School for Communication at the University of Pennsylvania, the Singapore Management University's Centre for AI and Data Governance, and the Tech Policy Design Centre at the Australian National University. CCG is a member of the Global Network Initiative (GNI), and our Executive Director, Jhalak M. Kakar, serves on the GNI Board as a member representing academics and academic organisations. She also serves on the Steering Committee for the Action Coalition on Meaningful Transparency, which aims to bring together a wide range of academics, civil society organisations, companies, governments, and international organisations in digital transparency.

The Centre has authored multiple publications over the years, including the two editions of our book on Privacy and the Indian Supreme Court, an essay series on Democracy in the Shadow of Big and Emerging Tech, a comprehensive report on Intermediary Liability in India, an edited volume of essays on Emerging Trends in Data Governance, a guide for Drafting Data Protection Legislation: A Study of Regional Frameworks in collaboration with the United Nations Development Programme, and most recently a collaborative report on Social Media Regulation in Sri Lanka, India and Bangladesh. It has also published reports from the three phases of the Blockchain Project conducted in collaboration with the Tech Policy Design Centre at the Australian National University, which maps the blockchain ecosystem in India and Australia.

Privacy and data protection have been focus areas for CCG since its inception, and the Centre has shaped discourse in this domain through research and analysis, policy inputs, capacity building, and related efforts. In 2020, the Centre launched the Privacy Law Library, a global database that tracks and summarises privacy jurisprudence emerging in courts across the world, in order to help researchers and other interested stakeholders learn more about privacy regulation and case law. The PLL currently covers 250+ cases from 20+ jurisdictions globally and also contains a High Court Privacy Tracker that tracks emerging High Court privacy jurisprudence in India.

CCG also has an online 'Teaching and Learning Resource' database for sharing research-oriented reading references on information technology law and policy. In recent times, the Centre has also offered Certificate and Diploma Courses on AI Law and Policy, Technology Law and Policy, and First Principles of Cybersecurity. These databases and courses are designed to help students, professionals, and academicians build capacity and ensure their nuanced engagement with the dynamic space of existing and emerging technology and cyberspace, their implications for society, and their regulation. Additionally, CCG organises an annual International Summer School in collaboration with the Hans Bredow Institute and the Faculty of Law at the University of Hamburg in collaboration with the UNESCO Chair on Freedom of Communication at the University of Hamburg, Institute for Technology and Society of Rio de Janeiro (ITS Rio) and the Global Network of Internet and Society Research on contemporary issues of information law and policy.

ABBREVIATIONS

API	Application Programming Interface
CSAM	Child Sexual Abuse Material
DMA	Digital Markets Act
DP	Data Privacy
DSA	Digital Services Act
DSC	Digital Service Coordinator
EBDS	European Board for Digital Services
EC	European Commission
ECHR	European Convention on Human Rights
ECNL	European Center for Not-for-profit Law
EDMO	European Digital Media Observatory
EDPS	European Data Protection Supervisor
EU	European Union
EU Charter	Charter of Fundamental Rights of the European Union
GDPR	General Data Protection Regulation
GIFCT	Global Internet Forum to Counter Terrorism

GNI	Global Network Initiative
IIB	Independent Intermediary Body
IRA	Russia's Internet Research Agency
LEAs	Law Enforcement Agencies
LGBTQIA+	Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, and Asexual, and other diverse gender and sexual identities not specifically encompassed in the abbreviation.
ML	Machine Learning
MSME	Micro, Small and Medium Enterprises
NCII	Non-consensual Intimate Image Abuse
NDAs	Non-Disclosure Agreements
OECD	The Organisation for Economic Co-operation and Development
PII	Personally Identifiable Information
T&Cs	Terms and Conditions
TMRC	Twitter Moderation Research Consortium
UNHRC	United Nations Human Rights Council
VLOPs	Very Large Online Platforms
VLOSEs	Very Large Online Search Engines

TABLE OF CONTENTS

INTRODUCTION

Objectives and Overview

The contemporary online information ecosystem is increasingly shaped by digital platform enterprises, commonly known as "social media platforms". While these platforms were once hailed for fostering connectivity and amplifying diverse voices, they have become focal points of concern for their role in fuelling misinformation, exacerbating social divisions, and failing to address societal harms. Calls for greater accountability have grown louder, particularly as evidence mounts of platforms' opaque decision-making processes and their uneven responses to online risks across different regions.

In response, the European Union's Digital Services Act (DSA) has emerged as one of the most ambitious efforts to regulate online platforms. At its core, the DSA seeks to establish transparency as a key pillar of platform governance, requiring companies to disclose how they moderate content, manage online advertising, and deploy algorithmic systems. It also introduces oversight mechanisms such as audits, risk assessments, and mandated researcher access to platform data. While designed for the European context, the DSA is already shaping global conversations on platform regulation, raising important questions about how its principles might influence regulatory frameworks elsewhere, particularly in the Global South.[1]

The extent of the DSA's global impact will depend on several factors. Some platforms may voluntarily implement their transparency measures beyond the EU to maintain consistency in their policies. However, it is unlikely that they will extend all obligations globally, particularly those that increase regulatory scrutiny or involve costly compliance measures. Instead, governments in other regions will have to determine whether and how they wish to adapt similar regulatory models, taking into account their unique political and economic conditions.

1 See for instance, Daphne Keller, 'The EU's new Digital Services Act and the Rest of the World' [2022] Verfassungsblog<https://verfassungsblog.de/dsa-rest-of-world/> accessed 15 October 2024.

For countries in the Global South, the prospect of platform transparency is particularly pressing, yet fraught with complexities. Many of these regions have already experienced disproportionate harm due to platform negligence—whether through failures in content moderation, algorithmic biases, or lack of investment in non-English language protections.[2] The impact of unchecked digital platforms has been especially severe in places where online hate speech and misinformation have translated into real-world violence, political instability, and the suppression of marginalised communities.[3] In such contexts, regulatory measures inspired by the DSA could provide crucial tools to hold platforms accountable and protect fundamental rights.

However, implementing meaningful transparency mechanisms in the Global South comes with significant challenges. Many governments lack the institutional capacity to enforce stringent disclosure requirements or oversee platform compliance effectively. Others operate in environments where transparency obligations could be misused—either to pressure platforms into content takedowns or to expand surveillance over online activity. Additionally, corporate incentives to comply with transparency regulations will differ across regions, with smaller economies often struggling to exert the same regulatory leverage as the EU.

---

2 See for instance, Cat Zakrzewski and others, 'How Facebook Neglected the Rest of the World, Fueling Hate Speech and Violence in India' *Washington Post* (24 October 2021) <https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/>; Gabriel Nicholas and Aliya Bhatia, 'Toward Better Automated Content Moderation in Low-Resource Languages' (2023) 2 Journal of Online Trust and Safety <https://www.tsjournal.org/index.php/jots/article/view/150> accessed 2 September 2024.

3 See for instance, Jasper Jackson, Mark Townsend and Lucy Kassa, 'Facebook "Lets Vigilantes in Ethiopia Incite Ethnic Killing"' *The Observer* (20 February 2022) <https://www.theguardian.com/technology/2022/feb/20/facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing> accessed 5 June 2023; Reuters, 'Myanmar: UN Blames Facebook for Spreading Hatred of Rohingya' *The Guardian* (13 March 2018) <https://www.theguardian.com/technology/2018/mar/13/myanmar-un-blames-facebook-for-spreading-hatred-of-rohingya> accessed 5 June 2023; Mike Isaac and Kevin Roose, 'Disinformation Spreads on WhatsApp Ahead of Brazilian Election' *The New York Times* (19 October 2018) <https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html> accessed 1 May 2021; Jeff Horwitz and Newley Purnell, 'YouTube, Facebook and Instagram Gave Platforms to Indian Cow-Protection Vigilante' *Wall Street Journal* (6 March 2023) <https://www.wsj.com/articles/youtube-facebook-and-instagram-gave-platforms-to-indian-cow-protection-vigilante-526833b6> accessed 7 March 2023.

Despite these obstacles, the DSA has set a new baseline for discussions on platform governance, one that governments, industry, civil society, and researchers worldwide will increasingly have to engage with. This report[4] examines the opportunities and risks associated with adapting DSA-inspired transparency measures in the Global South, recognising that effective regulation requires more than just legal mandates. It must be accompanied by robust institutional frameworks, strong civil society engagement, and safeguards against regulatory overreach, ensuring that transparency serves the objective of enhancing public accountability over the operations of online platforms.

Contextualising Social Media Platform Transparency: A Brief History

The last decade has witnessed a significant shift in discourses surrounding the social and political role of social media platforms. Celebrated widely in the immediate aftermath of the Arab Spring for facilitating vigorous civic and political engagement, social media's relationship with democracy is now under renewed and stricter scrutiny. At least since 2015, when reports first emerged regarding the profiling of Facebook users to influence electoral outcomes in the US,[5] frequent revelations, reports, and research findings have implicated popular social media platforms in a range of societal harms. Platforms have been associated with the rampant proliferation of misinformation and hate speech,[6] polarisation of public discourse,[7]

4 The DSA has been brought into force in a phased manner, and several delegated legislations associated with the transparency mechanisms outlined in this report are in different stages of deliberation, adoption and implementation. This report reflects the developments in legislation and implementation as of 31st October 2024

5 Harry Davies, 'Ted Cruz Using Firm That Harvested Data on Millions of Unwitting Facebook Users' *The Guardian* (11 December 2015) <https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data>.

6 See Isaac and Roose (n 3); David Klepper and Krutika Pathi, 'As India Votes, Misinformation Surges on Social Media: "The Whole Country Is Paying the Price"' *AP News* (2 May 2024) <https://apnews.com/article/india-election-misinformation-meta-youtube-703a56c73f9341393f05400ea218b87d>; David Gilbert, 'Hate Speech on Facebook Is Pushing Ethiopia Dangerously Close to a Genocide' *VICE* (14 September 2020) <https://www.vice.com/en/article/hate-speech-on-facebook-is-pushing-ethiopia-dangerously-close-to-a-genocide/>.

and large-scale violations of privacy and cybersecurity in numerous jurisdictions.[8] Challenging their early characterisation as amplifiers of marginalised voices, they have been found to have contributed to the stifling of dissent,[9] discrimination against minorities and the (re)production of hierarchies, including those of gender, race, caste and class,[10] and adverse mental health outcomes.[11] The breadth and scale of

---

7 See Cat Zakrzewski and others (n 2); Timothy McLaughlin, 'How Facebook's Rise Fueled Chaos and Confusion in Myanmar' *Wired*<https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/> accessed 1 April 2025; Michael Savage, 'How Brexit Party Won Euro Elections on Social Media – Simple, Negative Messages to Older Voters' *The Observer* (29 June 2019) <https://www.theguardian.com/politics/2019/jun/29/how-brexit-party-won-euro-elections-on-social-media> accessed 1 April 2025; Craig Timberg, Elizabeth Dwoskin and Reed Albergotti, 'Inside Facebook, Jan. 6 Violence Fueled Anger, Regret over Missed Warning Signs' *The Washington Post* (22 October 2021) <https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/>; Craig Timberg, 'Russian Propaganda Effort Helped Spread "Fake News" during Election, Experts Say' *The Washington Post* (25 November 2016)<https://www.washingtonpost.com/business/economy/russian-propaganda-effort-helped-spread-fake-news-during-election-experts-say/2016/11/24/793903b6-8a40-4ca9-b712-716af66098fe_story.html>; Reuters (n 3).

8 See 'EU Privacy Regulator Fines Meta 251 Million Euros for 2018 Breach' *Reuters* (17 December 2024) <https://www.reuters.com/technology/eu-privacy-regulator-fines-meta-251-million-euros-2024-12-17/> accessed 1 April 2025; Matthew Rosenberg, Nicholas Confessore and Carole Cadwalladr, 'How Trump Consultants Exploited the Facebook Data of Millions' *The New York Times* (17 March 2018) <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html> accessed 1 April 2025; 'Millions of Hacked LinkedIn IDs Advertised "for Sale"' *BBC News* (18 May 2016) <https://www.bbc.com/news/technology-36320322> accessed 1 April 2025; 'China: Weibo Admits to Leak of Personal Data on Millions of Users' (*Business & Human Rights Resource Centre*, 27 March 2020)<https://www.business-humanrights.org/en/latest-news/china-weibo-admits-to-leak-of-personal-data-on-millions-of-users/> accessed 1 April 2025; Reuters, 'Hackers Reportedly Leak Email Addresses of More than 200 Million Twitter Users' *The Guardian* (6 January 2023) <https://www.theguardian.com/technology/2023/jan/05/twitter-users-data-hacked-email-address-phone-numbers> accessed 1 April 2025.

9 See Deborah Brown and Rasha Younes, 'Meta's Broken Promises' (2023) <https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and>; Marwa Fatafta, 'How Meta Censors Palestinian Voices' (*Access Now*, 19 February 2024) <https://www.accessnow.org/publication/how-meta-censors-palestinian-voices/> accessed 1 April 2025.brown

10 See Shirin Ghaffary, 'The Algorithms That Detect Hate Speech Online Are Biased against Black People' (*Vox*, 15 August 2019) <https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter>; Taylor Lorenz, 'Instagram Is Full of Conspiracy Theories and Extremism' *The Atlantic* (21 March 2019) <https://www.theatlantic.com/technology/archive/2019/03/instagram-is-the-internets-new-home-for-hate/585382/>; Melany Amarikwa, 'Social Media Platforms' Reckoning: The Harmful Impact of TikTok's Algorithm on People of Color' (2023) 29 Richmond Journal of Law & Technology <https://www.ssrn.com/abstract=4349202> accessed 1 April 2025; 'Anti-Immigrant Violence Erupts across Britain amid Social Media Misinformation' *The Indian Express* (6 August 2024)

these harms amplify concerns surrounding the need for greater accountability from major platforms, for the risks that originate from their designs and their advertising-based business models.

While demands for accountability are articulated divergently by different stakeholders, a key thread running through them is their problematisation of the opacity with which platforms operate. As private actors, major platforms have historically guarded their design-choices, policies, practices and procedures from public view, citing the need to protect their intellectual property and their users' privacy. Consequently, there is limited information in the public domain on how they function, how they generate revenue, and how their algorithmic systems make decisions to shape the flow of information and content online. This information asymmetry reinforces their positions as *de facto* gatekeepers of online speech and expression. In this context, greater platform transparency, i.e. enhanced access to relevant information regarding platforms for various stakeholders, can be an important means towards building greater public accountability over platforms.

Enhanced transparency in the operation of social media platforms can deliver a range of beneficial outcomes for various audiences. For the billions of users, insights

---

<https://indianexpress.com/article/world/violence-erupts-across-britain-amid-misinformation-9497244/>.

11 See Brian A Primack and others, 'Use of Multiple Social Media Platforms and Symptoms of Depression and Anxiety: A Nationally-Representative Study among U.S. Young Adults' (2017) 69 Computers in Human Behavior 1 <https://www.sciencedirect.com/science/article/pii/S0747563216307543> accessed 1 April 2025; Sunkyung Yoon and others, 'Is Social Network Site Usage Related to Depression? A Meta-Analysis of Facebook–Depression Relations' (2019) 248 Journal of Affective Disorders 65 <https://www.sciencedirect.com/science/article/pii/S0165032718321700> accessed 1 April 2025; Elizabeth M Seabrook, Margaret L Kern and Nikki S Rickard, 'Social Networking Sites, Depression, and Anxiety: A Systematic Review' (2016) 3 JMIR Mental Health e50 <http://mental.jmir.org/2016/4/e50/> accessed 1 April 2025; Lucas Silva Lopes and others, 'Problematic Social Media Use and Its Relationship with Depression or Anxiety: A Systematic Review' (2022) 25 Cyberpsychology, Behavior and Social Networking 691; Betul Keles, Niall McCrae, and Annmarie Grealish, 'A Systematic Review: The Influence of Social Media on Depression, Anxiety and Psychological Distress in Adolescents' (2020) 25 International Journal of Adolescence and Youth 79 <https://doi.org/10.1080/02673843.2019.1590851> accessed 1 April 2025; Jean M Twenge and others, 'Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time' (2018) 6 Clinical Psychological Science 3 <https://doi.org/10.1177/2167702617723376> accessed 1 April 2025.

into platforms' designs and content-related decisions can enable informed decision-making and meaningful grievance redressal online. For researchers and academics, it can facilitate a deeper and more context-sensitive examination of the relationship between platforms and (say) civic discourse, mental health or gender-based violence. For public authorities, it can pave the way for evidence-based, principled policy measures to mitigate the risks posed by interactions on platforms. In these ways, platform transparency can result in greater trust and integrity in the social media information ecosystem.

Major platforms also increasingly recognise this. As research by Gorwa and Garton Ash outlines, platforms have attempted to offer more transparency to regain the trust of users, public authorities and civil society, particularly in response to public pressure following the 2016 US election.[12] In 2018, Facebook published detailed 'Community Standards', adding key contextual information for users on its content moderation policies. Shortly thereafter, it also released 'Enforcement Reports' setting out the various types of content takedowns it performed and aggregated data on the content it removed. Google and Twitter launched similar reports on the enforcement of their own terms and conditions. In parallel, platforms also initiated measures to enhance transparency in online political advertising, requiring advertiser verification and disclosing ad spending. Facebook created a public political ad archive, while Twitter and Google launched online transparency portals for ad-related disclosures.

In addition to these disclosures, platforms have also undergone external assessments by organisations or coalitions with which they have been affiliated. Such assessments, as discussed in further detail in Chapter III (*Risk Management*) and Chapter IV (*Audits*), have been important sources of information on platforms' operations. Further, as elaborated in Chapter V (*Researcher Access to Platform Data*), certain platforms have also provided access to some data for public interest research,

12 Robert Gorwa and Timothy Garton Ash, 'Democratic Transparency in the Platform Society', *Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge University Press 2020) <https://www.cambridge.org/core/books/social-media-and-democracy/democratic-transparency-in-the-platform-society/F4BC23D2109293FB4A8A6196F66D3E41> accessed 10 November 2023.

through Application Programming Interfaces (APIs) and partnerships with academic institutions.

These measures have doubtlessly thrown some light on platforms' functioning and contributed towards their comprehensibility for external stakeholders. At the same time, there are natural limitations to the efficacy of these measures. Fundamentally, their success, and often their continuation, hinges on platforms' largesse and/or continued cooperation with relevant third parties. While regulatory and political pressures and public relations can play a significant role in prompting and sustaining such initiatives, platforms retain the discretion to pull the plug without facing any legal consequences. Meta's recent closure of CrowdTangle, a tool used widely by researchers and journalists to monitor misinformation, illustrates this well.[13] Similar dependencies have undercut the promise of public APIs and research partnerships, with many major platforms unilaterally terminating them, limiting their scope or withholding relevant data in recent years, citing risks to privacy and security.[14] Moreover, it is near-impossible for external stakeholders to verify the reliability of the information disclosed by platforms, systematically draw comparisons between platforms, or contest platforms' claims regarding purported risks associated with sharing such data.

Given the consensus around the need for social media platform transparency, and the inherent limitations of platform-driven initiatives, it is no surprise that transparency mechanisms are increasingly finding a place in legislative proposals for platform regulation. As discussed in each chapter of this report, these mechanisms

13 Sarah Grevy Gotfredsen and Kaitlyn Dowling, 'Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn't as Transparent or Accessible' (*Columbia Journalism Review*, 9 July 2024) <https://www.cjr.org/tow_center/meta-is-getting-rid-of-crowdtangle.php>.

14 See Axel Bruns, 'After the "APIcalypse": Social Media Platforms and Their Fight against Critical Scholarly Research' (2019) 22 Information, Communication & Society 1544 <https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1637447> accessed 11 April 2024; Douglas Parry, 'Restrictions on Data Access Impede Crucial Societal Research' (*University World News*, 9 May 24AD) <https://www.universityworldnews.com/post.php?story=20240509000643443>; Sasha Moriniere, 'We Must Fix Researcher Access to Data Held by Social Media Platforms' (*Canvas*, 16 November 2023) <https://medium.com/odi-research/we-must-fix-researcher-access-to-data-held-by-platforms-9084de211854>.

often build on platforms' existing voluntary initiatives, reinforcing them through legal enforcement. Germany's Network Enforcement Act (or NetzDG),[15] brought into force in 2018, was perhaps the first legislation to impose periodic transparency reporting requirements on major platforms. Shortly thereafter, Canada (Bill C-76),[16] and France (Law No. 2018-1202)[17] passed legislation requiring platforms to present information on advertisements hosted by them.

The Digital Services Act: A New Chapter in Platform Transparency

The intention to introduce the European Union's Digital Services Act (DSA) was first articulated by Ursula von der Leyen in 2019, in her agenda for presidency of the European Commission (EC).[18] The DSA, designed to update EU's extant regulations on online trust and safety, was envisioned as a part of a package of EU regulations for the "digital age". The EC officially proposed the initial version of the DSA in December 2020 alongside the Digital Markets Act (DMA), designed to address the gatekeeping power of major platforms. After extensive negotiations, the European Parliament and EU member states reached a political agreement on the DSA in April 2022.[19] It was finally passed into law in October 2022,[20] followed closely by the DMA

15Netzwerkdurchsetzungsgesetz (NetzDG) 2017 <https://perma.cc/7UCW-AA3A>. >

16 Elections Modernization Act 2018, c 31 <https://www.parl.ca/DocumentViewer/en/42-1/bill/C-76/royal-assent>.

17 See Irène Couzigou, 'The French Legislation Against Digital Information Manipulation in Electoral Campaigns: A Scope Limited by Freedom of Expression' (2021) 20 Election Law Journal: Rules, Politics, and Policy 98 <https://www.liebertpub.com/doi/10.1089/elj.2021.0001> accessed 1 April 2025; Nicolas Boring, 'Initiatives to Counter Fake News: France' (*Library of Congress Law*, April 2019) <https://maint.loc.gov/law/help/fake-news/france.php> accessed 1 April 2025.

18 Ursula von der Leyen, *Political Guidelines for the Next European Commission 2019–2024* (European Commission 2019) <https://commission.europa.eu/document/download/063d44e9-04ed-4033-acf9-639ecb187e87_en?filename=political-guidelines-next-commission_en.pdf>.

19 European Commission, 'Digital Services Act: Commission welcomes political agreement on rules ensuring a safe and accountable online environment' (Press Release IP/22/2545, 23 April 2022) < https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545>.

in November 2022.[21] By this time, laws requiring disclosure regarding aspects of major platforms' operations had been passed in many jurisdictions, including Australia[22], India,[23] and the state of California in the US.[24] Nonetheless, the DSA is one of the most comprehensive legislative interventions on platform transparency.

Recognising the role of online intermediaries in shaping the exercise of rights guaranteed under the Charter of Fundamental Rights of the EU, including the right to freedom of expression and information, the DSA aims to ensure their responsible behaviour for a safe, predictable, and trustworthy online environment. It applies to a range of providers of digital "intermediary services", where such services are offered to recipients established or located in the EU.[25] It lays down a tiered categorisation of intermediary services, which include hosting services, which in turn include "online platforms".[26] Further, online platforms and online search engines with more than 45 million active users in the EU, on average each month, are designated as Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs).[27] As of October 2024, 17 online platforms were designated as VLOPs under the DSA.[28] This report focuses on online platforms, including Very Large Online Platforms (VLOPs).

---

20 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065>.

21 Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)<https://eur-lex.europa.eu/eli/reg/2022/1925/oj/eng>.

22 Australian Online Safety Act of 2021<https://www.legislation.gov.au/C2021A00076/latest/text>.

23 IT (intermediary Guidelines and Digital Media Ethics Code) Rules 2021 <https://cbcindia.gov.in/wp-content/uploads/2021/09/IT-book-English.pdf>.

24California Assembly Bill No. 587, <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB587>.

25 DSA 2022, art 2(1).

26 DSA 2022, recital 41; DSA art 3; See 'The EU's Digital Services Act' (*European Commission*, 27 October 2022) <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en>.

27 DSA 2022, art 33.

28 'Supervision of the Designated Very Large Online Platforms and Search Engines under DSA' (European Commission) <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>.

Carrying forward the legacy of the EU's e-Commerce Directive,[29] the DSA retains the prohibition on the imposition of any general obligation on intermediaries to monitor user-generated content.[30] It also preserves the "safe harbour" rule, conditionally protecting intermediaries from liability for such content.[31] Further, it leaves member-states with the discretion to determine the legality of online content under national laws.[32] Instead, it focuses on intermediaries' procedural conduct and their decision-making in relation to online content, setting out an extensive range of due diligence obligations, which are independent from the determination of intermediaries' liability for online content.[33] These are intended towards ensuring the safety and trust of users (particularly those most vulnerable to online harms), fundamental rights under the EU Charter and meaningful accountability of online intermediaries.

The DSA treats promoting the transparency of online services as a core objective towards building their accountability. Accordingly, its due diligence framework includes a range of obligations under which intermediaries must disclose information regarding their services to various sets of stakeholders, including the EC, national authorities, researchers, auditors, users and the general public. These obligations are tailored in accordance with the type, size, and nature of the intermediary – some apply to all intermediaries, others apply only to hosting service providers, while some others apply specifically to online platforms. The most stringent set of obligations apply only to VLOPs and VLOSEs, in view of their popularity, scale, and significance in the online information ecosystem.[34] The Appendix to this report maps and summarises the transparency obligations under the DSA, along with their applicability, periodicity and audience. These relate to

29 Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') < https://eur-lex.europa.eu/eli/dir/2000/31/oj/eng>.

30 DSA 2022, art 8. DSA 2022, recital 30.

31 DSA 2022, art 1(a); DSA 2022, chapter II; DSA 2022, recitals 16-28.

32 DSA 2022, art 3(h); DSA 2022, recital 12.

33 DSA 2022, recital 41; Also see Martin Husovec, 'Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules' [2023] SSRN Electronic Journal <https://www.ssrn.com/abstract=4598426>.

34 DSA 2022, recital 41.

various aspects of intermediaries' services, including their terms and conditions, content moderation procedures, recommender systems and advertising practices.

In addition to disclosure obligations, the DSA also requires VLOPs and VLOSEs to submit to certain oversight mechanisms, which are expected to reveal more information regarding their services. VLOPs and VLOSEs must assess the potential systemic risks posed by their services every year, and implement measures to mitigate the identified risks.[35] Further, they must commission annual audits, where external auditors assess their compliance with their obligations under the DSA and recommend operational measures to ensure compliance.[36] Each of these procedures must be documented and publicly reported every year, subject to countervailing considerations of confidentiality, user privacy and public security.[37] Moreover, VLOPs and VLOSEs must provide researchers access to data to facilitate research regarding the detection, identification and understanding of systemic risks posed by their services.[38] Such research can be expected to shed further light on the impact of VLOPs' and VLOSEs' services in specific jurisdictional contexts, particularly on vulnerable groups. Cumulatively, the disclosure obligations and oversight mechanisms prescribed under the DSA promise to offer valuable insights into intermediaries' systems and operations for external stakeholders and to contribute to a sharper public understanding of intermediaries' services and their societal impact.

For its implementation and oversight, the DSA requires each member-state to designate a Digital Services Coordinator (DSC).[39] It also sets up the European Board for Digital Services (EBDS), an independent advisory group of DSCs to supervise the application and evolution of the DSA.[40] Departing from the division of powers between national authorities and the EC under the EU's General Data Protection Regulation (GDPR), the DSA reserves many enforcement powers exclusively for the

35 DSA 2022, arts 34-35.

36 DSA 2022, art 37.

37 DSA 2022, art 42(4).

38 DSA 2022, art 40.

39 DSA 2022, art 49.

40 DSA 2022, art 61.

EC. This can be expected to mitigate uneven implementation across EU member-states, and to avoid disproportionate enforcement burden on regulators in states housing the headquarters of many VLOPs and VLOSEs. The regulatory costs incurred by the EC for the implementation of the DSA will be discharged through the annual supervisory fees levied on VLOPs and VLOSEs.[41] The operationalisation of many such obligations will be supported by delegated acts,[42] voluntary standards,[43] and codes of conduct under the DSA.[44] The failure of an intermediary to comply with obligations under the DSA, including its transparency obligations, can attract penalties up to 6% of their annual worldwide turnover.[45]

Given the vast institutional framework required for its implementation, the DSA has been brought into force in a phased manner, with all provisions becoming applicable on November 17, 2024.

While it is too early to assess the success of the DSA in enhancing meaningful transparency, it has already resulted in the disclosure of swathes of hitherto unavailable information regarding major platforms' operations. As of October 2024, many VLOPs and VLOSEs have published two sets of their biannual transparency reports, revealing crucial patterns in their content moderation practices.[46] The DSA Transparency Database, aggregating platforms' statements of reasons for their content moderation decisions, contains close to 10 billion statements by more than 80 platforms.[47] The first cycle of risk assessments, risk mitigation, external audits and audit implementation of VLOPs and VLOSEs is also complete.[48]

---

41 DSA 2022, art 43; DSA 2022, recital 101.

42 DSA 2022, art 86.

43 DSA 2022, art 44.

44 DSA 2022, arts 45-47.

45 DSA 2022, art 52.

46 Tremau T&S Research Team, 'DSA Database' (*Tremau*) <https://tremau.com/resources/dsa-database>.

47 EC, 'Digital Services Act Transparency Database' <https://transparency.dsa.ec.europa.eu/> accessed 16 December 2023.

48 Tremau T&S Research Team (n 46).

The Global Impact of the DSA

The DSA is a landmark legislation in platform regulation, and its impact is likely to extend beyond the EU's borders, especially given the bloc's power to unilaterally set the agenda for global regulation through what has been termed the "Brussels effect".[49]

The extraterritorial impact of the DSA remains to be seen and is likely to vary across jurisdictions. Regulatory convergence in platform regulation can occur either *de facto* with platforms voluntarily adopting the obligations under the DSA globally or *de jure* through regulatory imitation by other jurisdictions.[50]

Voluntary implementation by platforms (*de facto*) depends on economic incentives to adopt uniform policies and processes globally, as well as technical feasibility for compliance to be localised.[51] De facto convergence in some areas might be easier than in others. For instance, the extraterritorial impact of DSA-mandated content moderation, notice and takedown, and grievance redressal systems on the future governance of online speech is being carefully observed.[52]

When it comes to de facto standardisation of transparency obligations, certain obligations like more detailed Terms and Conditions (T&Cs) and explanations for recommender systems might be globally extended.[53] However, platforms are not

49 Anu Bradford, 'The Brussels Effect' (2012) 107 Northwestern University Law Review 1 <https://heinonline.org/HOL/P?h=hein.journals/illlr107&i=1> accessed 1 May 2023.

50 ibid.

51 Martin Husovec and Jennifer Urban, 'Will the DSA Have the Brussels Effect?' [2024] Verfassungsblog<https://verfassungsblog.de/will-the-dsa-have-the-brussels-effect/> accessed 14 October 2024.

52 Dawn C Nunziato, 'The Digital Services Act and the Brussels Effect on Platform Content Moderation' 24 Chicago Journal of International Law; Husovec and Urban (n 51); Daphne Keller, 'The Rise of the Compliant Speech Platform' (*Lawfare*, 16 October 2024) <https://www.lawfaremedia.org/article/the-rise-of-the-compliant-speech-platform> accessed 29 October 2024.

53 Keller (n 1); Laureline Lemoine and Mathias Vermeulen, 'The Extraterritorial Implications of the Digital Services Act - DSA Observatory' (1 November 2023) <https://dsa-observatory.eu/2023/11/01/the-extraterritorial-implications-of-the-digital-services-act/> accessed 15 October 2024.

likely to extend many other transparency obligations to jurisdictions beyond the EU unless legally mandated. Obligations like risk assessments, data access or audits will be costly and expose platforms to additional public and regulatory scrutiny without providing any short-term gains.[54] Already, many major platforms maintain different compliance standards for different jurisdictions,[55] and it might be easy to continue to do so for transparency obligations under the DSA. Thus, even when platforms maintain voluntary disclosures globally, providing additional information mandated by the DSA may not be extended to all jurisdictions. For instance, voluntary ad archives maintained by several platforms have only made targeting information available for ads displayed in the EU.[56]

Regulatory imitation (*de jure*) across different jurisdictions in the Global South will depend on local factors, including existing regulatory frameworks, legislative priorities, regulatory costs, political structures, and the presence of strong civil society, academia, and media. Further, market power often determines what resources platforms allocate for content moderation and regulatory compliance across different jurisdictions, and it might be difficult for smaller countries to impose obligations similar to those of the DSA.[57] Since the EU is a single market, it gives the union considerable leverage to impose novel and higher regulatory standards across member states. However, geopolitical factors in many regions prevent similar economic integration (see South Asia)[58] further diminishing the bargaining power of smaller countries.

---

54 See Husovec and Urban (n 51).

55 ibid.

56 See 'Ad Library API' <https://www.facebook.com/ads/library/api/?source=nav-header> accessed 30 October 2024; 'Ads Transparency' <https://support.google.com/adspolicy/answer/13733850?hl=en&ref_topic=13775718&sjid=9562857601950222877-AP> accessed 30 October 2024.

57 See Zahra Takhshid, 'Regulating Social Media in the Global South' (2021) 24 Vanderbilt Journal of Entertainment & Technology Law 1 <https://scholarship.law.vanderbilt.edu/jetlaw/vol24/iss1/1>; Giovanni De Gregorio and Nicole Stremlau, 'Inequalities and Content Moderation' (2023) 14 Global Policy 870 <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.13243> accessed 7 February 2024.

58 Santosh Sharma Poudel, 'SAARC Is Dead. Long Live Subregional Cooperation' (*The Diplomat*, 27 September 2022) <https://thediplomat.com/2022/09/saarc-is-dead-long-live-sub-regional-co-operation/>.

Even when similar transparency obligations are contemplated in the Global South, the way the provisions are transposed and operationalised might be very different from how they are envisaged in the EU. The DSA requires strong societal structures, like communities of researchers, empowered civil society actors, and a specialised and independent regulator to ensure effective implementation and meaningful accountability.[59] For instance, transparency mechanisms like data access for public interest research provide an immense opportunity to examine the online information ecosystem and platforms' moderation and curation of user-generated content and advertisements. However, factors such as a diminished environment for independent research, limited data protection laws, and restricted funding and infrastructure might impede the ability of many Global South jurisdictions to operationalise researcher access to platform data.

In fact, in some jurisdictions, there might also be the potential risk of transparency mechanisms being appropriated for direct and collateral surveillance and censorship.[60] Further, the creation of a regulator truly independent from the executive, like the DSC, might be difficult in the immediate term for many Global South jurisdictions.[61] In several jurisdictions, backdoor negotiations between platforms and states might result in transparency provisions becoming a bureaucratic compliance exercise or a tool to jawbone[62] platforms to alter their content moderation practices.

All this is not to say that the DSA might not result in similar risks in the EU. After all, the EC, tasked exclusively with supervising and enforcing the DSA against VLOPs and VLOSEs, is not an independent regulator like the DSCs, but an executive body

59 Martin Husovec, 'Will the DSA Work?' [2022] Verfassungsblog<https://verfassungsblog.de/dsa-money-effort/> accessed 14 August 2024.

60 See Anupam Chander, 'When the Digital Services Act Goes Global' [2023] Georgetown Law Faculty Publications and Other Works <https://scholarship.law.georgetown.edu/facpub/2548>.

61 See Keller (n 1); Chander (n 60).

62 Jawboning refers to "scenarios where the government pressures online services to moderate user content in ways going beyond their legal authority to do so." See P.J. Leerssen, 'Jawboning and lawboning: comparing platform-state relations in the US and Europe' (*CELE,* 2024) <https://pure.uva.nl/ws/files/218376920/CELE_post.pdf>.

guided by the dominant political agenda of the Union.[63] The EC's attempts at enforcement of the DSA have already prompted due process concerns. In his letters to major platforms in the immediate aftermath of the escalation of violence between Israel and Palestine in October 2023, Thierry Breton (then European Commissioner for Internal Market) conflated illegal content and "disinformation" – a concept that, as many civil society organisations noted, evades definition under human rights law.[64] The letters also focused on swift content removals and requested platforms to be in contact with law enforcement agencies and "respond promptly to their requests".[65] This interpretation of the DSA could potentially result in suppressing the flow of crucial information during severe conflict, including coverage of human rights violations in Gaza.[66] This echoes concerns that the EC or member states could potentially use provisions in the DSA to censor legitimate speech – both by targeting specific pieces of content and by influencing platforms' designs and policies.[67] As Griffin notes, such concerns have become especially pertinent with the rise of far-right political parties in Europe in recent years, and their increasing representation in the European Parliament.[68]

---

63 Ilaria Buri, 'A Regulator Caught Between Conflicting Policy Objectives' (*Verfassungsblog*, 31 October 2022) <https://verfassungsblog.de/dsa-conflicts-commission/> accessed 30 October 2024.

64 See 'Precise Interpretation of the DSA Matters Especially When People's Lives Are at Risk in Gaza and Israel' (*Access Now*, 18 October 2023) <https://www.accessnow.org/press-release/precise-interpretation-of-dsa-matters-in-gaza-and-israel/> accessed 20 October 2024; 'Europe: Tackling Content about Gaza and Israel Must Respect Rule of Law' (*ARTICLE 19*, 18 October 2023) <https://www.article19.org/resources/europe-tackling-content-gaza-israel-must-respect-rule-of-law/> accessed 29 October 2024; Itxaso Domínguez de Olazábal and others, 'Position Paper on Palestinian Digital Rights and the Extraterritorial Impact of the European Union's Digital Services Act (DSA)' <https://7amleh.org/storage/Advocacy%20Reports/English%207amleh.pdf>.

65 See Letter from Breton to Musk (10 October 2023) <https://x.com/ThierryBreton/status/1711808891757944866>; Letter from Breton to Zuckerberg (10 October 2023) <https://x.com/thierrybreton/status/1712126600873931150?s=46&t=Y-CDvNYEVAdPCdphstKDgQ>.

66 'Europe: Tackling Content about Gaza and Israel Must Respect Rule of Law' (n 64); de Olazábal and others (n 64); 'Precise Interpretation of the DSA Matters Especially When People's Lives Are at Risk in Gaza and Israel' (n 64).

67 Rachel Griffin, 'EU Platform Regulation in the Age of Neo-Illiberalism' (Social Science Research Network, 29 March 2024) <https://papers.ssrn.com/abstract=4777875> accessed 18 October 2024.

68 ibid; Federica Marsi, 'European Parliament at Crossroads as Right-Wing Parties Triumph in EU Vote' *Al Jazeera* (12 June 2024) <https://www.aljazeera.com/features/2024/6/12/european-parliament-at-crossroads-as-right-wing-parties-triumph-in-eu-vote> accessed 27 November 2024;

The DSA is an outcome of political negotiations in the backdrop of a unique history of platform regulation in Europe. Similarly, other jurisdictions have their own unique political, historical, and economic contexts, and unilateral regulatory convergence might not be easily achievable or even desirable. However, this must not be understood to say that jurisdictions work in isolation from one another. It is undeniable that the DSA has set the discourse around platform transparency in contemporary times, and almost all future discussions on platform regulation and accountability will have to contend with the new baseline set by it.

As Keller notes, the DSA is, in some ways, an "experiment" with many first-time obligations like risk assessments, audits, data access for public research, and user notification for visibility reductions.[69] Civil society and regulators across the globe are keeping a close eye on the developments in Europe. The DSA can have an indirect impact on content moderation in what Hussovec and Urban refer to as the "shipping container moment, giving an entire industry vocabulary, structure and building blocks".[70] For instance, standardisation of content moderation might result in some level of convergence in transparency reporting over a period of time.[71] Similarly, the guidelines laid down by the DSA for audits and risk assessments are, in all likelihood, going to provide a reference point for all future discussions on similar frameworks in other jurisdictions.

The discussion on transparency obligations in the DSA becomes especially crucial for the Global South. Many countries in this region are grappling with the proliferation of child sexual abuse material (CSAM), non-consensual intimate images (NCII), violent and extremist content, hate speech, and disinformation online. Whistleblower reports have revealed significant lapses by platforms, causing severe

---

Jennifer Rankin, 'Germany's AfD and Extremist Allies Set up Second EU Parliament Far-Right Group' *The Guardian* (11 July 2024) <https://www.theguardian.com/world/article/2024/jul/11/germany-afd-extremist-allies-set-up-second-eu-parliament-far-right-group> accessed 27 November 2024.

69 Keller (n 1).

70Husovec and Urban (n 51).

71 ibid.

violations of human rights in the Global South.[72] Content moderation in non-English languages, spoken by the majority of users on the Internet, has been insufficient.[73] The risks are significantly exacerbated for low-resource languages where platforms operating on a profit motive are not incentivised to invest in building automated tools or employing moderators embedded in the local linguistic and cultural context.[74] The scale of negligence of non-English languages can be gauged from whistleblower revelations on Facebook spending as much as "87% of its resources on tackling misinformation in English when only 9% of its users are English speakers".[75] Content moderation is also frequently outsourced to contractual workers, many of whom reside in the Global South and work under abysmal conditions.[76]

All this has contributed to endangering some of the most vulnerable and persecuted communities in the Global South. In India, hate speech against minority groups has

72 Craig Silverman, Ryan Mac, and Pranav Dixit, '"I Have Blood On My Hands": A Whistleblower Says Facebook Ignored Global Political Manipulation' *BuzzFeed News* (14 September 2020) <https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo> accessed 29 October 2024; Vittoria Elliott and others, 'The Facebook Papers Reveal Staggering Failures in the Global South' (*Rest of World*, 26 October 2021) <https://restofworld.org/2021/facebook-papers-reveal-staggering-failures-in-global-south/> accessed 7 May 2024; Caroline Crystal, 'Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech' (*Carnegie Endowment for International Peace*, 7 September 2023) <https://carnegieendowment.org/research/2023/09/facebook-telegram-and-the-ongoing-struggle-against-online-hate-speech?lang=en> accessed 30 October 2024.

73 'How Big Tech Platforms Are Neglecting Their Non-English Language Users' (*Global Witness*, 30 November 2023) <https:///en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/> accessed 30 October 2024; De Gregorio and Stremlau (n 57).

74 See Nicholas and Bhatia (n 2); De Gregorio and Stremlau (n 57); Mona Elswah, 'Moderating Maghrebi Arabic Content on Social Media' (30 September 2024) <https://osf.io/3n849> accessed 29 October 2024; Gabriel Nicholas, 'The Dire Defect of "Multilingual" AI Content Moderation' *Wired*<https://www.wired.com/story/content-moderation-language-artificial-intelligence/> accessed 2 September 2024.

75 Dan Milmo, 'Facebook Revelations: What Is in Cache of Internal Documents?' *The Guardian* (25 October 2021) <https://www.theguardian.com/technology/2021/oct/25/facebook-revelations-from-misinformation-to-mental-health> accessed 30 October 2024.

76 Tarleton Gillespie, 'The Human Labor of Moderation', *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (Yale University Press 2018); '"The Despair and Darkness of People Will Get to You"' (*Rest of World*, 22 July 2020) <https://restofworld.org/2020/facebook-international-content-moderators/> accessed 3 September 2024.

often led to violence[77] and Brazil has grappled with disinformation during elections, culminating in violent rioting.[78] Independent investigations have uncovered how inflammatory content on Facebook contributed to ethnic massacres in war-torn Ethiopia.[79]Facebook's role in the dehumanisation of Rohingya Muslims in Myanmar, culminating in ethnic cleansing in 2017, was established by the UN fact-finding mission.[80]

Users in the Global South, especially those belonging to historically marginalised communities, often face the double brunt of online hate speech and targeted harassment on one hand and censorship of legitimate expression and dissent on the other. Meaningful platform accountability is an important step to safeguard the rights of the most vulnerable communities. Transparency in the interaction between platforms and the state is also crucial to protect freedom of expression online.

While it remains to be seen how the DSA's transparency provisions will be operationalised, many jurisdictions in the Global South are rethinking platform regulation, and the DSA is likely to become an important point of reference. In this report, we examine the opportunities and challenges that the DSA's transparency provisions could present for jurisdictions in the Global South. We acknowledge that meaningful transparency requires a multi-faceted approach, entailing both independent and effective governance structures as well as, a dynamic and empowered civil society to hold states and platforms accountable. Legal frameworks mandating platform transparency must be supported by complementary regulations on data protection, competition and intellectual property. Simultaneously, legislative

77 'Facebook Failing to Check Hate Speech, Fake News in India: Report' *Al Jazeera* (25 October 2021) <https://www.aljazeera.com/news/2021/10/25/facebook-india-hate-speech-misinformation-muslims-social-media> accessed 30 October 2024; Horwitz and Purnell (n 3); Billy Perrigo, 'Facebook Let an Islamophobic Conspiracy Theory Flourish in India Despite Employees' Warnings' [2021] *Time*<https://time.com/6112549/facebook-india-islamophobia-love-jihad/>.

78 Elizabeth Dwoskin, 'Come to the "War Cry Party": How Social Media Helped Drive Mayhem in Brazil' *Washington Post* (9 January 2023) <https://www.washingtonpost.com/technology/2023/01/08/brazil-bolsanaro-twitter-facebook/> accessed 30 October 2024.

79 Jackson, Townsend and Kassa (n 3).

80 Reuters (n 3).

reforms must go hand-in-hand with efforts to strengthen justice delivery systems, corporate accountability structures, and oversight of private power by independent media and civil society. Only through these parallel processes working in tandem can long-term changes be expected in how platforms are designed and operated.

METHODOLOGY

At the outset, we conducted a comprehensive textual analysis of the Digital Services Act (DSA) to identify and examine the transparency-related obligations imposed on online intermediaries. Our objective was to explore how the DSA operationalises transparency and the ways in which it mandates intermediaries to disclose information. This includes direct disclosure obligations as well as procedural requirements such as risk assessments and audits that facilitate information-sharing. Each provision mandating transparency has been systematically tabulated and summarised in the Appendix detailing the categories of intermediaries to which the provision applies, the frequency of disclosure, and the intended recipients or beneficiaries. Additionally, to enhance accessibility for researchers navigating the DSA's transparency framework, we have thematically classified the provisions and applied a structured color-coded system The transparency mechanisms in Annex I are classified into the following seven categories: (i) transparency in recommending; (ii) transparency in advertising; (iii) risk management; (iv) audits; (v) researcher access to platform data; (vi) transparency in content moderation; (vii) disclosures to regulatory authorities (other than those covered under the mechanisms above).

Our research was guided by a multi-pronged methodological approach, incorporating primary and secondary research methods as well as expert consultations and interactive discussions. This approach enabled us to build a robust understanding of the transparency mechanisms embedded in the DSA and their implications for platform regulation, particularly in the Global South.[1]

1 The DSA has been brought into force in a phased manner, and several delegated legislations associated with the transparency mechanisms outlined in this report are in different stages of deliberation, adoption and implementation. This report reflects the developments in legislation and implementation as of 31st October 2024

1.Legal Text and Policy Analysis

We began our study with a close textual analysis of the DSA to extract transparency obligations imposed on intermediaries. This legal examination was supplemented by an analysis of the European Commission's documentation, legislative debates, and official guidance to understand the intended scope and objectives of these transparency mechanisms.

Our analysis of the provisions in the Appendix led to the identification of six key transparency mechanisms under the DSA that enhance public understanding of platform policies, procedures, and practices.[2] These include:

- Transparency in platform recommendation systems
- Transparency in advertising practices
- Systemic risk management, including risk assessment, mitigation, and reporting
- External audits for compliance with due diligence obligations
- Data access mechanisms for public-interest research
- Transparency in content moderation, including aggregated reports, user notifications, and Terms and Conditions (T&Cs) disclosures

We recognised that while these mechanisms share a common overarching goal — enhancing transparency — they differ significantly in scope, purpose, procedural requirements, and the stakeholders they target.

2 These comprise of any transparency measure that culminates in information disclosure to the public. The report does not examine platform disclosures to state authorities for the purpose of compliance and enforcement of the DSA or any other law. For instance, data access for public interest research, might lead to initial data disclosures to researchers, but ultimately, the insights from public-interest research is expected to reach audiences beyond academia. Similarly, the findings of the audit reports and risk assessments are meant to be made publicly available.

2. Literature Review

To contextualise our findings, we conducted an extensive literature review of interdisciplinary scholarship on social media platform governance, transparency, and regulatory frameworks. This included:

- Academic research on social media platform regulation, transparency, and content moderation.
- Policy reports and legal analyses of the DSA, including official EU documentation.
- Public consultations and stakeholder submissions made to the European Commission on the DSA.
- Research on the political economy of social media platforms, particularly in Global South jurisdictions.

3. Expert Interviews and Stakeholder Consultations

To deepen our understanding of both the EU and Global South regulatory contexts, we conducted semi-structured interviews with experts working on platform regulation, including those affiliated with the Global Network Initiative. Our interviewees included:

- Researchers and legal scholars studying the DSA and social media regulation in the EU.
- Legal practitioners and policy analysts engaged in platform governance in the Global South, including experts from Brazil, South Korea, Sri Lanka, Singapore, and India.

These interviews provided valuable qualitative insights into the implementation challenges and practical efficacy of transparency mechanisms, especially in varying political, legal, and economic contexts.

4. Interactive Feedback and Expert Workshops

To refine our findings and incorporate expert feedback, we actively participated in policy dialogues and research workshops on platform regulation. Notably, our preliminary findings were discussed and enriched through:

- The "DSA and Platform Regulation Conference 2024" organised by the DSA Observatory at the Institute for Information Law (IViR), Amsterdam Law School, University of Amsterdam (February 2024).
- The Data Governance Network Quarterly Roundtable in Mumbai (August 2023)
- Workshops hosted by the Global Network Initiative, where we engaged with global experts on digital rights and platform accountability.
- The roundtable titled "Behind the Glassdoor: Social Media Transparency through the Indian Lens" (August 2024), organised by the Centre for Communication Governance (CCG).

These engagements allowed us to validate our research findings and integrate cross-disciplinary perspectives into our analysis.

5. Contextual Analysis: Implications for the Global South

Recognising that regulatory approaches cannot be transplanted wholesale, we examined the feasibility of adapting the DSA's transparency mechanisms to Global South jurisdictions. Our assessment considered:

- Socioeconomic conditions, including digital literacy levels and internet penetration.
- Political and regulatory structures, such as the presence (or absence) of independent regulators.

- Legal and enforcement challenges, including jurisdictional constraints over multinational platforms.

- The role of civil society, particularly the extent to which civil society actors can demand accountability from platforms as well as states.

This analysis was grounded in our institutional expertise in India, as well as broader research on the Global South's digital governance landscape. While we acknowledge the diversity of Global South jurisdictions, our study provides a foundational framework for further region-specific research on platform transparency.

6. Limitations of Our Study

We recognise that many of the challenges highlighted in this report for Global South jurisdictions may also apply to the EU. However, our focus remains on identifying challenges and opportunities specific to Global South contexts.

Furthermore, the "Global South" is a highly contested, dynamic and diverse category.[3] It is not monolithic, and includes highly diverse jurisdictions across Asia, Africa, Latin America, the Middle East, and Oceania. Political structures range from liberal democracies to autocracies, regulatory capacities vary widely, and economic conditions shape platform governance differently in each region. While our study attempts to account for these complexities, we acknowledge that our perspective is influenced by our location and expertise in India, and may not capture the full spectrum of regional nuances.

3 Sebastian Haug, Jacqueline Braveboy-Wagner and Günther Maihold, 'The "Global South" in the Study of World Politics: Examining a Meta Category' (2021) 42 Third World Quarterly 1923 <https://doi.org/10.1080/01436597.2021.1948831> accessed 5 November 2024; Stewart Patrick and Alexandra Huggins, 'The Term "Global South" Is Surging. It Should Be Retired.' (*Carnegie Endowment for International Peace*, 15 August 2023) <https://carnegieendowment.org/posts/2023/08/the-term-global-south-is-surging-it-should-be-retired?lang=en> accessed 5 November 2024.

Conclusion

By integrating legal analysis, interdisciplinary literature review, expert interviews, policy workshops, and contextual analysis, our methodology ensures a well-rounded, evidence-based understanding of the DSA's transparency framework. We hope that this study serves as a valuable resource for scholars, policymakers, and civil society actors engaged in the ongoing global discourse on platform transparency and regulation.

Brief Outline of the Report

In the subsequent chapters of this report, we analyse each transparency mechanism separately while also noting important interlinkages between them where relevant. Each chapter begins with a brief introduction outlining the mechanism, its potential value in enhancing transparency and the historical context in which it was incorporated in the DSA. This is followed by a deeper analysis of the mechanism, highlighting the transparency gains that it can be expected to achieve for various stakeholders and its limitations. Each chapter pays special attention to the suitability of the mechanism for adaptation in Global South jurisdictions and concludes with our insights on the challenges that Global South jurisdictions attempting such adaptation must navigate. Our insights on each chapter are compiled in the final chapter of the report, titled, "Key Insights for the Global South", for ease of reference.

1. TRANSPARENCY IN RECOMMENDING CONTENT

1.1. Introduction

A defining feature of contemporary online platforms, and in particular, social media platforms, is their ability to curate content for their users and to influence the relative priority in which it is presented. Platforms perform this task ("recommending"), using recommender systems – that typically involve the use of machine-learning algorithms that gauge users' preferences and accordingly deliver customised (and increasingly personalised) feeds of content to them.[1]

Recommender systems are important to the optimal functioning of the online 'marketplace of content'. On the demand-side, they enable content-recipients to navigate through the swathes of posted content, mitigating information-overload and filtering for relevance. On the supply-side, they assist content-publishers in reaching audiences that are likelier to engage with their content, and advertisers, in targeting audiences that are likelier to consume the goods and services they advertise.[2] Put differently, not only does a recommender system determine the nature of content likely to be seen by a particular user, but also the reach of a particular piece of content to users.[3]

1 *See* Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' (2019) 10(3) EJLT <https://ejlt.org/index.php/ejlt/article/view/686> accessed November 10, 2024. The authors categorise recommender systems into two technical categories: those that perform 'content-based filtering', recommending content based on similarity to content previously consumed by the user; and those that perform 'collaborative filtering', recommending content based on the content similar users have consumed. A third 'hybrid' category combines the features of both.

2 For a more detailed analysis of the role of recommender systems in online advertising, *see* Chapter II (*Ad transparency*).

3 Arvind Narayanan, 'Understanding Social Media Recommendation Algorithms' (Knight First Amendment Institute 2023) <http://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms> accessed 22 August 2023.

Given their central role in determining the flow of information via online platforms, recommender systems are seen as tools through which platforms exercise the power to shape online experiences, and influence public discourse and user-behaviour.[4] However, unlike gatekeepers of traditional mass media (which have been constrained over time, at least in some measure, by journalistic and broadcasting standards), recommender systems are shaped primarily by platforms' economic incentive to optimise for users' engagement.[5] Thus, a key objective of most recommender systems is to algorithmically rank content for each user according to the systems' prediction of *how likely the user is to engage with such content*.[6] In systems with personalised output, this likelihood is predicted by processing a wide range of weighted signals and data-inputs, including the user's demographic information, information on the user's networks and/or the user's online behaviour, including their chosen subscriptions.[7]

As research has increasingly revealed, this quest for engagement-optimisation poses significant risks, at both the individual and the societal levels. These include creating 'echo chambers'[8] or 'filter bubbles',[9] propagating disinformation,[10] hate speech[11] and

4 Silvia Milano, Mariarosaria Taddeo and Luciano Floridi, 'Recommender Systems and Their Ethical Challenges' (2020) 35 AI & SOCIETY 957 <https://doi.org/10.1007/s00146-020-00950-y> accessed 22 August 2023; Christian Djeffal, Christina Hitrova and Eduardo Magrani, 'Recommender Systems and Autonomy: A Role for Regulation of Design, Rights, and Transparency' (2021) 17 Indian Journal of Law and Technology <https://www.ijlt.in/_files/ugd/066049_6ef9b3b0dad94943a20865b42c5d138c.pdf> accessed 22 August 2023.

5 Paddy Leerssen, 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems' (24 February 2020) <https://papers.ssrn.com/abstract=3544009> accessed 22 August 2023; Milano and others (n 4); Djeffal and others (n 4);

6 Narayanan (n 3); Cobbe and Singh (n 1).

7 *See* Narayanan (n 3), which describes three stylised models of information propagation on social media platforms: namely, subscription-based models, network-based models and algorithm-based models; Oliver Budzinski and Madlen Karg, 'Algorithmic Search and Recommender Systems in the Digital Services Act' [2022] Competition Policy International <https://www.competitionpolicyinternational.com/wp-content/uploads/2022/12/6-ALGORITHMIC-SEARCH-AND-RECOMMENDER-SYSTEMS-IN-THE-DIGITAL-SERVICES-ACT-By-Oliver-Budzinski-Madlen-Karg.pdf>accessed 22 August 2023.

8 *See* Noémi Bontridder and Yves Poullet, 'The Role of Artificial Intelligence in Disinformation' (2021) 3 Data & Policy e32 <https://www.cambridge.org/core/journals/data-and-policy/article/role-of-artificial-intelligence-in-disinformation/7C4BF6CA35184F149143DE968FC4C3B6> accessed 22 August 2023.

conspiracy theories,[12] 'digital gerrymandering',[13] amplifying dominant narratives,[14] and stifling dissenting ones.[15]

Notably, certain studies have also contested the purportedly widespread nature of echo chambers and filter bubbles. *See* Amy Ross Arguedas and others, 'Echo chambers, filter bubbles, and polarisation: a literature review' (Reuters Institute 2022), <https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review> accessed 12 April 2024; Andrew Guess and others, AVOIDING THE ECHO CHAMBER ABOUT ECHO CHAMBERS: Why selective exposure to like-minded political news is less prevalent than you think' (Knight Foundation 2018), <https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf> accessed 12 April 2024.

9 Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan, 'Exploring the Filter Bubble: the effect of using recommender systems on content diversity' (2014) In Proceedings of the 23rd International Conference on World Wide Web <https://dl.acm.org/doi/abs/10.1145/2566486.2568012> accessed 22 August 2023; Bontridder and Poullet (n 8).

10 Miriam Fernandez and Alejandro Bellogín, 'Recommender Systems and Misinformation: The Problem or the Solution?'; Joe Whittaker and others, 'Recommender Systems and the Amplification of Extremist Content' (2021) 10 Internet Policy Review <https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content> accessed 22 August 2023; Bontridder and Poullet (n 8); Christian Stöcker, 'How Facebook and Google Accidentally Created a Perfect Ecosystem for Targeted Disinformation' in Christian Grimme and others (eds), Disinformation in Open Online Media (Springer International Publishing 2020); Hannah Ellis-Petersen, 'Revealed: Meta approved political ads in India that incited violence' *The Guardian* (20 May 2024) <https://www.theguardian.com/world/article/2024/may/20/revealed-meta-approved-political-ads-in-india-that-incited-violence> accessed 9 September 2024; .

11 Sahana Udupa and Matti Pohjonen, 'Extreme Speech| Extreme Speech and Global Digital Cultures — Introduction' (2019) 13 International Journal of Communication 19 <https://ijoc.org/index.php/ijoc/article/view/9102> accessed 22 August 2023; 'Letting hate flourish: YouTube and Koo's lax response to the reporting of hate speech against women in India and the US', Global Witness and Internet Freedom Foundation (February 2024) <https://www.globalwitness.org/en/campaigns/digital-threats/letting-hate-flourish-youtube-and-koos-lax-response-to-the-reporting-of-hate-speech-against-women-in-india-and-the-us/> accessed 9 September 2024; Koh Ewe and Cape Diamond, 'Myanmar Coup: Soldiers Flood TikTok With Calls to Violence' *VICE* (3 March 2021) <https://www.vice.com/en/article/myanmar-coup-soldiers-flood-tiktok-with-calls-to-violence/> accessed 9 September, 2024.

12 Elise Thomas, 'Recommended Reading: Amazon's Algorithms, Conspiracy Theories and Extremist Literature' (ISD 2021) <https://www.isdglobal.org/isd-publications/recommended-reading-amazons-algorithms-conspiracy-theories-and-extremist-literature/> accessed 22 August 2023; Megan A. Brown and others, 'Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users' (SSRN Electronic Journal 2022) <https://ssrn.com/abstract=4114905> accessed 22 August 2023; Muhsin Yesilada and Stephan Lewandowsky, 'Systematic Review: YouTube Recommendations and Problematic Content' (2022) 11 Internet Policy Review <https://policyreview.info/articles/analysis/systematic-review-youtube-recommendations-and-problematic-content>accessed 22 August 2023,

Despite overwhelming consensus regarding the existence of these risks, it has been difficult to correlate particular harms, arising in particular contexts, with the functioning of particular recommender systems, due to critical information-gaps. Citing trade secret protection and risks of manipulation by malicious actors,[16] platforms have closely guarded any information on how they define relevance and user-engagement, which data-inputs they leverage, which parameters they use to promote or demote content and the relative significance of these parameters.[17] In rare cases, where platforms have disclosed certain components of recommendation algorithms in response to public pressure, they have kept other crucial information

---

13 Jonathan Zittrain, 'Engineering an Election' (2014) 127 Harvard Law Review Forum <https://harvardlawreview.org/forum/vol-127/engineering-an-election/> accessed 28 August 2023.

14 Michael Färber, Melissa Coutinho and Shuzhou Yuan, 'Biases in Scholarly Recommender Systems: Impact, Prevalence, and Mitigation' (2023) 128 Scientometrics 2703 <https://doi.org/10.1007/s11192-023-04636-2> accessed 22 August 2023.

15 Working Group on Pluralism of News and Information in Curated and Indexing Algorithms, 'Pluralism of News and Information in Curated and Indexing Algorithms: Policy Framework' (Forum on Information & Democracy 2023) <https://informationdemocracy.org/wp-content/uploads/2023/02/Report_Pluralism-in-algorithms.pdf> accessed 22 August 2023; 'Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook', Human Rights Watch (21 December 2023) <https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and> accessed 12 April 2024; Marwa Fatafta, 'It's not a glitch: how Meta systematically censors Palestinian voices', *AccessNow* (19 February 2024) <https://www.accessnow.org/publication/how-meta-censors-palestinian-voices/> accessed 9 September 2024; Paige Collins and Jillian C. York, 'Digital Apartheid in Gaza: Unjust Content Moderation at the Request of Israel's Cyber Unit' Electronic Frontier Foundation (26 July 2024) <https://www.eff.org/deeplinks/2024/07/digital-apartheid-gaza-unjust-content-moderation-request-israels-cyber-unit> accessed 25 January 2025;'Civil society to Meta: Stop censoring reproductive rights content' *AccessNow* (1 September 2022) <https://www.accessnow.org/press-release/meta-stop-censoring-reproductive-rights-content/> accessed 9 September 2024.

16 Robin K Hill, 'Gaming the System: Definition' (CACM, 31 July 2021) <https://cacm.acm.org/blogs/blog-cacm/254472-gaming-the-system-definition/fulltext?mobile=false> accessed 22 August 2023; Leerssen (n 5); Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society <https://doi.org/10.1177/2053951715622512> accessed 22 August 2023.

17 Burrell (n 16); Amélie Heldt, Matthias C Kettemann and Paddy Leerssen, 'The Sorrows of Scraping for Science: Why Platforms Struggle with Ensuring Data Access for Academics' (*Verfassungsblog*, 30 November 2020) <https://verfassungsblog.de/the-sorrows-of-scraping-for-science/> accessed 22 August 2023;Paddy Leerssen, 'Algorithm Centrism in the DSA's Regulation of Recommender Systems' (*Verfassungsblog*, 29 March 2022) <https://verfassungsblog.de/roa-algorithm-centrism-in-the-dsa/> accessed 22 August 2023.

concealed inside the 'black-box'.[18] The unavailability of such information has precluded a holistic understanding of social media recommender systems and the ways in which they may act as vectors of online harms. Thus, as a first step towards holding platforms accountable for the risks they create, and towards effective mitigation, more information on their recommender systems is key.

In one of the first *regulatory* efforts towards recommender systems transparency, the DSA attempts to throw light on recommender systems. Casting a wide net in defining them, it covers fully and partially automated systems, as well as systems that recommend in response to search-prompts. Intermediaries using such systems must disclose to users the main parameters they use in recommending content.[19] They must present this information in their terms and conditions in an easily comprehensible manner. Alongside, they must provide reasons for the relative importance assigned to the parameters.[20] Further, where an intermediary's recommender system allows for the selection and modification of such parameters, the intermediary must provide users a directly-accessible functionality to do so, in its online interface.[21]

Additionally, VLOPs and VLOSEs using recommender systems must provide users at least one option, which is not based on profiling.[22] In effect, this obligates such VLOPs and VLOSEs to offer users an option to view content without any personalisation – illustratively, by choosing to receive content ranked simply in chronological order, or to receive content that is shared by other users in their

18 Frank Pasquale, *The Block Box Society*: *The Secret Algorithms That Control Money and Information* (Harvard University Press 2015) <https://raley.english.ucsb.edu/wp-content/Engl800/Pasquale-blackbox.pdf> accessed 18 November 2023; Arvind Narayanan, 'Twitter Showed Us Its Algorithm. What Does It Tell Us?' (Algorithmic Amplification and Society, Knight First Amendment Institute), 10 April 2023.

19 DSA 2022, art 27(1).

20DSA 2022, art 27(2).

21DSA 2022, art 27(3).

22 DSA 2022, art 38; See GDPR, art 4(4), which defines 'profiling' as any form of automated processing of personal data involving its use to evaluate certain personal aspects relating to a person, in particular to analyse or predict aspects concerning their performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

network.[23] In addition to these user-facing obligations, they must explain to state authorities, upon request, the design, the logic, the functioning and the testing of such systems.[24]

1.2. Disclosure of and control over parameters

The DSA recognises algorithm-based recommendations as a "core part" of the platform's businesses, which has significant implications on users' access to information and online behaviour.[25]

As research has demonstrated, many users are currently unaware of even the use of algorithmic systems by platforms to deliver content to them.[26] Thus, instead of recognising the content-feed presented to them as driven by the platform's estimation of their interests, values and networks, many users perceive it as a random or chronological assortment, or, as a neutral representation of reality.[27] This may partly explain why users across jurisdictions have increasingly started relying on social media platforms for news and information, as recent surveys indicate.[28]

---

23 To comply with this requirement, Meta modified its systems to allow EU-based users an option to view Stories and Reels only from people they follow, ranked in chronological order. See Nick Clegg, 'New Features and Additional Transparency Measures as the Digital Services Act Comes Into Effect' (*Meta*, 22 August 2023) <https://about.fb.com/news/2023/08/new-features-and-additional-transparency-measures-as-the-digital-services-act-comes-into-effect/> accessed 28 January 2025.

24DSA 2022, art 40(3).

25 DSA 2022, recital 70.

26 Spandana Singh, 'Rising Through the Ranks: How Algorithms Rank and Curate Content in Search Results and on News Feeds' (Open Technology Institute 2019) <https://www.newamerica.org/oti/reports/rising-through-ranks/> accessed 22 August 2023; Motahhare Eslami and others, 'FeedVis: A Path for Exploring News Feed Curation Algorithms', Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (ACM 2015) <https://dl.acm.org/doi/10.1145/2685553.2702690> accessed 22 August 2023.

27 Spandana Singh (n 26).

28 Pew Research Center, 'Social Media and News Fact Sheet' (Pew Research Centre 2022) <https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/> accessed 14 November 2023; Oxford University Press, The Matter of Fact (OUP 2022) <https://oup.foleon.com/the-matter-of-fact/the-matter-of-fact-2022/> accessed 14 November 2023.

This holds particularly true in Global South jurisdictions, where digital literacy and access to digital services remain fragmented, despite the rapid expansion of technological infrastructures in the last decade.[29] These conditions accentuate concerns, often expressed through the lenses of "algorithmic literacy" and "the algorithmic divide", regarding the patchy understanding of the role played by recommender systems in delivering content.[30] This leaves users in such jurisdictions especially vulnerable to online manipulation. In recent years, coordinated manipulation or similar targeted behaviour driven on social media has called into question the integrity of national elections in Brazil, Philippines and Kenya,[31] and as per numerous reports, fuelled violence against minority groups in India, Myanmar and Ethiopia.[32]

---

29 See International Telecommunication Union, Measuring digital development: Facts and Figures: Focus on Least Developed Countries (ITU 2023) <https://www.itu.int/pub/D-IND-ICT_MDD> accessed 14 November 2023; International Telecommunication Union, Measuring digital development: Facts and Figures: Focus on Least Developed Countries (ITU 2022) <https://www.itu.int/hub/publication/d-ind-ict_mdd-2022/> accessed 14 November 2023; Telecom Regulatory Authority of India, Consultation Paper No. 16/2023, on Digital Inclusion in the Era of Emerging Technologies (TRAI 2023) <https://www.trai.gov.in/consultation-paper-digital-inclusion-era-emerging-technologies> accessed 14 November 2023.

30 Anne Oeldorf-Hirsch and German Neubaum, 'What do we know about algorithmic literacy? The status quo and a research agenda for a growing field' (2023) New Media & Society <https://doi.org/10.1177/14614448231182662> accessed 14 November 2024; Peter K. Yo, 'The algorithmic divide and equality in the age of artificial intelligence' (2020) 72 Florida Law Review 331.

31 Odango Madung, 'Brazil, Kenya, the US – tech giants are putting democracy in peril the world over' Guardian (London, 25 Jan 2023) <https://www.theguardian.com/global-development/2023/jan/25/brazil-kenya-the-us-tech-giants-are-putting-democracy-in-peril-the-world-over> accessed 15 November 2023; Wilson Centre, 'Episode 17: The Impact of Misinformation on Brazil's Elections' (October 27 2022) <https://www.wilsoncenter.org/microsite/2/node/112277> accessed 15 November 2023; Gerargo Eusebio, 'Fake news and internet propaganda, and the Philippine elections: 2022' Rappler (23 May 2022) <https://www.rappler.com/technology/features/analysis-fake-news-internet-propaganda-2022-philippine-elections/> accessed 25 January 2025..

32 Maya Mirchandani, 'Digital Hatred, Real Violence: Majoritarian Radicalisation and Social Media in India (Observer Research Foundation, August 2018) <https://www.orfonline.org/research/43665-digital-hatred-real-violence-majoritarian-radicalisation-and-social-media-in-india/> accessed 15 November 2023. Caroline Crystal, 'Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech' Carnegie Endowment for International Peace (7 Sept 2023) <https://carnegieendowment.org/2023/09/07/facebook-telegram-and-ongoing-struggle-against-online-hate-speech-pub-90468> accessed 15 November 2023; Billy Perrigo, 'Facebook Was Used to Incite Violence in Myanmar. A New Report on Hate Speech Shows It Hasn't Learned Enough Since Then' Time (London, 29 Oct 2019) <https://time.com/5712366/facebook-hate-speech-violence/>

In this context, clear disclosure of the use of recommender systems, the parameters used by such systems and their relative significance, can facilitate at least an elementary understanding for a user as to why content appears to them in the given manner and priority. Drawing from van Drunen *et al*,[33] such disclosures in content personalisation enhance users' trust in platforms in three interrelated ways. First, they allow users to align their expectations with the platform's objectives, as gleaned from the disclosures.[34] Where a platform discloses to a user that an article on COVID-19 has been recommended due to its endorsement by jurisdictional health authorities, the user can estimate that the platform strives to support the government's awareness-initiatives surrounding the pandemic. They can accordingly navigate through the intermediary's platform reasonably expecting that on matters of public health, content that aligns with government-led initiatives would be prioritised. Second, such disclosures enable users to channel scepticism, preventing the adoption of a generalised scepticism to online content.[35] Where the user is reasonably aware of the platform's prioritisation of content that aligns with their government's public health initiatives, they can exercise critical judgment over content relating to public health. Instead of indiscriminately accepting (or rejecting) the information therein, they can assess it against advisories issued by the World Health Organisation or academic articles in public health journals. Third, such disclosures facilitate trust-repair, by allowing platforms to demonstrably modify their recommender systems to respond to perceived trust-violations.[36] Where amplification of the voices of prominent political leaders has persistently contributed to misinformation on a platform during a public health crisis, the platform can

---

accessed 15 November 2023; Amnesty International, 'Ethiopia: Meta's failures contributed to abuses against Tigrayan community during conflict in northern Ethiopia' (31 October 2023), <https://www.amnesty.org/en/latest/news/2023/10/meta-failure-contributed-to-abuses-against-tigray-ethiopia/> accessed 9 September 2024.

33 Max van Drunen, Brahim Zarouali and Natali Helberger, 'Recommenders You Can Rely on: A Legal and Empirical Perspective on the Transparency and Control Individuals Require to Trust News Personalisation' (2022) 13 Journal of Intellectual Property, Information Technology and Electronic Commerce Law <https://www.jipitec.eu/issues/jipitec-13-3-2022/5562> accessed 22 August 2023.

34 van Drunen and others (n 33).

35 ibid.

36 ibid.

change its recommender systems to actively foster scientifically informed perspectives.

While the DSA is geared towards these objectives, scholars have pointed to weaknesses in its attempt to provide meaningful control for users over content recommended by platforms. While Article 27 notionally provides users the ability to modify the parameters of a recommender system where several options are available, it does not oblige platforms to offer such options.[37] Without a legal obligation, VLOPs and VLOSEs are particularly unlikely to offer such options, considering the immense market-power that they wield in many jurisdictions.

Another weakness pointed out widely by scholars is that the DSA does not require intermediaries to take the user's affirmative consent to 'profile' them in order to recommend content, instead assuming such consent by default.[38] In fact, except

37 ARTICLE 19, 'Digital Services Act: ARTICLE 19 Proposed Amendment to Article 29 Recommender Systems' <https://www.article19.org/wp-content/uploads/2021/05/Amendment-recommender-systems.pdf> accessed 22 August 2023; Huw Roberts and others, 'Governing Artificial Intelligence in China and the European Union: Comparing Aims and Promoting Ethical Outcomes' (2023) 39 The Information Society <https://www.tandfonline.com/doi/full/10.1080/01972243.2022.2124565>accessed 22 August 2023;Natali Helberger and others, 'Regulation of news recommenders in the Digital Services Act: Empowering David against the very large online Goliath." (2021) 26 Internet Policy Review <https://policyreview.info/articles/news/regulation-news-recommenders-digital-services-act-empowering-david-against-very-large> accessed 22 August 2023; Djeffal and others (n 4); Maximilian Gahntz, 'Towards Responsible Recommending: Recommendations for Policymakers & Large Online Platforms v1.0' (Mozilla 2022) <https://assets.mofoprod.net/network/documents/Mozilla_Towards-Responsible-Recommending.pdf> accessed 22 August 2023; European Data Protection Supervisor (EDPS), 'Opinion 1/2021 on the Proposal for a Digital Services Act <https://edps.europa.eu/system/files/2021-02/21-02-10-opinion_on_digital_services_act_en.pdf> accessed 22 August 2023; Fernandez and Bellogín (n 10).

38 Cobbe and Singh (n 1); Ilaria Buri and Joris van Hoboken, 'The Digital Services Act (DSA) Proposal: A Critical Overview' (Digital Services Act (DSA) Observatory, Institute for Information Law, University of Amsterdam 2021) <https://dsa-observatory.eu/wp-content/uploads/2021/11/Buri-Van-Hoboken-DSA-discussion-paper-Version-28_10_21.pdf> accessed 22 August 2023; ; Amnesty International, 'Amnesty International Position on the Proposals for a Digital Services Act and a Digital Markets Act' (March 2021) <https://www.amnesty.eu/wp-content/uploads/2021/04/Amnesty-International-Position-Paper-Digital-Services-Act-Package_March2021_Updated.pdf> accessed 22 August 2023; Miguel Pereira, 'Taming Europe's Digital Landscape? Brief Notes on the Proposal for a Digital Services Act' (2021) 7 UNIO – EU Law Journal 77 <https://revistas.uminho.pt/index.php/unio/article/view/4031> accessed 22 August 2023; Matúš Mesarčík and others, 'Analysis of Selected Regulations Proposed by the European Commission and Technological Solutions in Relation to the Dissemination of Disinformation and the Behaviour of Online Platforms.' (Kempelen Institute of Intelligent Technologies 2022) <https://kinit.sk/wp-

VLOPs and VLOSEs, intermediaries can continue to make such profiling an essential condition for the delivery of recommended content to a user. Further, it frames the choice between personalisation and non-personalisation as a binary, thereby failing to secure granular control for users over their data.[39] For instance, a user of an e-commerce platform is not guaranteed the ability to allow the use of (say) their shopping history for personalised recommendations, but disallow the use of their other financial records or health records.

Pertinently, certain concerns relating to transparency and autonomy in the processing of personal data can be addressed by way of data privacy ('**DP**') laws. Such laws provide controls and place safeguards on the collection and use of personal data by any entity.[40] A rights-based DP framework would, at least, secure for users the rights to effectively access, modify and demand erasure of any personal data used by an intermediary towards recommending content.[41], It would provide users remedies against the unauthorised use of their data, through an empowered statutory authority. Further, it would entail mechanisms to ensure that companies adhere to responsible data practices and implement corrective measures, when algorithmic processing leads to unfair or otherwise harmful outcomes.

In fact, the DSA recognises the EU's General Data Protection Regulation (GDPR)[42] as one of its principal starting-points, and its protections only add to the gamut of rights

---

content/uploads/2022/04/Mesarcik-2022-Analysis-regulations-disinfoEN.pdf> accessed 22 August 2023.

39 Djeffal and others (n 4); Gahntz (n 37); EDPS (n 37).

40 Djeffal and others (n 4).

41 As eminent privacy scholars have pointed out, individual DP rights may not be sufficient to guarantee users control over their personal data. This may require broader structural measures that empower individuals and communities to make informed decisions on the use of their data, and promote organisational accountability over such use. *See* Daniel Solove, 'The Limitations of Privacy Rights' 98 (2023) Notre Dame Law Review 975 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4024790> accessed 16 November 2023.

42 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

guaranteed under the GDPR.[43] Further, the European Data Protection Supervisor (EDPS), and national data protection authorities in the EU, are expected to continue guiding the interpretation, evolution and enforcement of the DSA.[44] However, while DP laws are being increasingly adopted across jurisdictions, many Global South jurisdictions are yet to even propose a DP law.[45] Thus, at present, users in such jurisdictions do not even have a legal right to know of the personal data already collected or being collected to recommend content to them. Even where DP laws aiming to protect basic individual data rights have been enacted, they are marred by major lacunae, both substantively as well as in relation to their enforcement.[46]

---

43 DSA 2022, recital 10.

44 European Data Protection Supervisor, Shaping a Safer Digital Future: The EDPS Strategy 2020-24 (EDPS 2020) <https://edps.europa.eu/press-publications/publications/strategy/shaping-safer-digital-future_en> accessed 16 November 2023; European Data Protection Supervisor, Annual Report 2022 (EDPS 2023) <https://edps.europa.eu/2022-edps-annual-report/en/> accessed 16 November 2023.

45 *See* Graham Greenleaf, 'Global Data Privacy Laws 2023: 162 National Laws and 20 Bills' (2023) 181 Privacy Laws and Business International Report (PLBIR) 1, 2-4 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4426146> accessed 15 November 2023., setting out a list of jurisdictions that do not have a DP law. This includes Myanmar and Iraq, where major social media platforms have significant user-bases and where whistleblowers have revealed their algorithms' contributions to violent conflicts. *See* Mark Scott, 'Facebook did little to moderate posts in the world's most violent countries' (Politico 25 Oct 2021) <https://www.politico.eu/article/facebook-content-moderation-posts-wars-afghanistan-middle-east-arabic/> accessed 15 November 2023; Jeremy B. Merrill and Will Oremus, 'Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation' (The Washington Post, 26 Oct 2021) <https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/> accessed 15 November 2023.

46 For instance, in India, where a DP law has been enacted after numerous proposals, and Bangladesh, where a proposal is tabled, the statutory authorities' independence from the executive and their enforcement powers are, by design, limited. See Graham Greenleaf, 'India's 2023 Data Privacy Act: Business/government Friendly, Consumer Hostile' (2023) 185 Privacy Laws & Business International Report 1, 3-12 <https://ssrn.com/abstract=4666389> accessed 12 April 2024; AccessNow, 'India's Data Protection Bill a threat to privacy', (3 August 2023) <https://www.accessnow.org/press-release/indias-data-protection-bill/> accessed 18 November 2023; Graham Greenleaf, 'Bangladesh's Data Protection Bill' (2022) 179 Privacy Laws & Business International Report 26 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4343782> accessed 16 November 2023; Transparency International Bangladesh, 'The Revised Draft Data Protection Act (DPA) 2023: Review and recommendations in light of submissions on the earlier version' (2023) <https://www.ti-bangladesh.org/upload/files/position-paper/2023/Position-Paper-on-Revised-Draft-Data-Protection-Act-Review-Recommendations.pdf> accessed 16 November 2023.

The accumulation and deployment of personal data lie at the heart of personalised recommendations. Without basic levels of transparency and control for users in the processing of such data, the utility of specialised disclosure requirements aimed at parameters of recommender systems remains limited. Unless users have effective remedies against the collection and use of their personal data, they would have no means to challenge its use by an intermediary, even if they are informed of the parameters the intermediary uses to recommend content. Thus, Global South jurisdictions contemplating transparency frameworks for recommender systems must, on priority, adopt DP laws to undergird them.

1.3. Illuminating regulatory blind-spots

a. Looking beyond algorithmic parameters

The DSA creditably attempts to secure more information and provide choices to users on the main parameters used by platforms for recommending. Nevertheless, the centrality assigned to such parameters in its understanding of recommender systems merits closer scrutiny.

As noted earlier, the DSA's definition covers both fully and partially automated systems, as well as systems that respond to search-prompts.[47] The breadth allows the law to apply to a range of systems deployed by platforms, which vary significantly in their design.[48] However, Article 27's focus on the "main parameters" and their "relative importance" betrays a deficient understanding of recommender systems.

First, its vague formulation, while seemingly intuitive, does not correspond to the design of contemporary machine-learning algorithms that recommend content.[49] As

47 DSA 2022, art 3(s).

48 Cobbe and Singh conceptualise three non-technical forms of recommending, based on sources of information leveraged by platforms in recommending: 'open recommending' (used by Meta and Reddit), 'curated recommending' (used by Netflix) and 'closed recommending' (used by traditional news outlets); *see* Cobbe and Singh (n 1);

49 Jonathan Stray and others, 'Building Human Values into Recommender Systems: An Interdisciplinary Synthesis' (arXiv, 20 July 2022) <http://arxiv.org/abs/2207.10192> accessed 22

media researchers have noted, the "complexity"[50] of such algorithms, along with the scale at which they are deployed, renders them unsuitable to simplistic "causal explanations", even by technical experts.[51] In other words, it is impossible to conclusively link a specific outcome delivered by such algorithms (say, a particular piece of content recommended to a user), to particular parameter(s) that the algorithms employ. To understand how such algorithms propagate (or suppress) content, it is important to understand the way the algorithms are *configured*. including which engagement signals they use as inputs and how they define user-engagement.[52] The DSA does not concretely require the disclosure of any of these aspects.[53] Thus, its requirement to disclose the "main parameters" is only expected to provide a generic overview of how content is propagated (or suppressed) by a recommender system, much like Twitter's highly-publicised release of its "source code".[54]

August 2023; Helberger and others (n 37); Djeffal and others (n 4); Amnesty International (n 38); ARTICLE 19 (n 37) ; Buri and van Hoboken (n 38).

50 Describing recommender systems, Narayanan uses the term "complex" to refer to a system "whose behavior arises in nonlinear, often unpredictable ways from those of its parts."; *see* Narayanan (n 3).

51 Burrell (n 16); Leerssen (n 5); Budzinski and Karg (n 7); Lilian Edwards & Michael Veale (2017), 'Slave to the algorithm? Why a 'Right to an Explanation' is probably not the remedy you are looking for', 2017 Duke Law & Technology Review 16 <https://scholarship.law.duke.edu/dltr/vol16/iss1/2/> accessed 25 January 2025.

52 See Narayanan (n 3), which highlights that platforms vary significantly in the form of user-engagement that they optimise for. For instance, Meta reportedly optimises for "Meaningful Social Interactions", a weighted average that considers Likes, Reacts, Shares and Comments, while TikTok seems to optimise for a combination of Likes, Comments and the play-time of videos by users. Further, platforms often tweak their algorithms to change their optimisation-targets.

53 The DSA does leave scope for further specification of the contents of such disclosures through voluntary standards under Article 44(1)(e). However, platforms are unlikely to commit voluntarily to disclosing more than what is statutorily prescribed, in relation to their recommending systems.

54 Arvind Narayanan, 'Twitter Showed Us Its Algorithm. What Does It Tell Us?' (Algorithmic Amplification and Society, Knight First Amendment Institute, 10 April 2023) <https://knightcolumbia.org/blog/twitter-showed-us-its-algorithm-what-does-it-tell-us> accessed 18 November 2023. Sheila Dang, 'Twitter makes some of its source code public, promises more' (Reuters, 1 April 2023) <https://www.reuters.com/technology/twitter-makes-content-recommendation-code-public-2023-03-31/> accessed 18 November 2023; Paddy Leerssen, 'Outside the Black Box: From Algorithmic Transparency to Platform Observability in the Digital Services Act' (Weizenbaum Journal, 2024) <https://doi.org/10.34669/wi.wjds/4.2.3> accessed September 9, 2024.

But more importantly, disproportionate attention on algorithmic parameters ignores the fundamentally socio-technical nature of recommender systems.[55] As Leerssen notes, the algorithmic output of such systems, i.e. the content recommended to users, is not determined solely by the algorithm – it is also influenced by users in crucial ways.[56] Users themselves publish content on an intermediary's platform, which is then recommended to other users.[57] Further, user-behaviour on the platform, alongside other data-inputs relating to users, provides feedback signals – these influence the relative priority of algorithmic parameters over time, *via* the recursive process of machine-learning.[58] For instance, by watching Instagram reels of an Australian culinary show, a user in India can (perhaps unwittingly) nudge the algorithmic system to recommend itineraries for the Great Barrier Reef. Considering the role of user-behaviour in shaping the output of recommender systems, scholars have advocated for a shift of focus away from 'the algorithm', towards its mutually-influential interactions with users.[59] Some commentators have highlighted the significance of more details regarding the sources of user-data and the behavioural signals leveraged by a platform for recommending.[60] Others have argued for greater visibility over the algorithmic output, i.e. the content actually recommended to users over time, as well as how users actually engaged with such content.[61] Notably, the

55 Bernhard Rieder and Jeanette Hofmann, 'Towards platform observability' (2020) Internet Policy Review 9(4) <https://doi.org/10.14763/2020.4.1535> accessed 9 September 2024; Leerssen (n 17); Leerssen (n 5).

56 Leerssen (n 5).

57 Leerssen (n 5).

58 Leerssen (n 5).

59 Leerssen (n 17); Leerssen (n 5); Budzinski and Karg (n 7); Narayanan (n 3).

60 Jonathan Stray, 'Show me the algorithm: Transparency in recommendation systems' Schwartz Reisman Institute of Technology and Society (August 25, 2021) <https://srinstitute.utoronto.ca/news/recommendation-systems-transparency> accessed 25 January 2025; 'Fixing Recommender Systems' (Panoptykon Foundation, 2023) <https://panoptykon.org/sites/default/files/2023-08/Panoptykon_ICCL_PvsBT_Fixing-recommender-systems_Aug%202023.pdf> accessed 25 January 2025.

61 Rieder and Hofmann (n 55)., 'Towards platform observability' (2020) Internet Policy Review 9(4) <https://doi.org/10.14763/2020.4.1535> accessed 9 September 2024; Leerssen (n 54); Leerssen (n 5); Leerssen (n 17); Philip M Napoli, 'Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers' (2015) 39 Telecommunications Policy 751 <https://www.sciencedirect.com/science/article/pii/S030859611400192X> accessed 22

DSA does oblige VLOPs and VLOSEs to create publicly accessible repositories of advertisements shown to users for a year after which they are last shown,[62] as detailed in Chapter II (*Transparency in advertising*) – however, no such record is required for non-commercial content. Besides interactions with users, scholars have also highlighted the significance of organisational processes and personnel that steer the development and operation of recommender systems. These include the designers who develop such systems and the reviewers who supervise their functioning.[63] In this regard, the DSA does require VLOPs and VLOSEs to specify the human resources dedicated to content moderation in their T&Cs,[64] and transparency reports.[65] However, given that "content moderation" under the DSA relates only to activities aimed at illegal content or content incompatible with T&Cs, [66] such disclosures may not reveal the human interventions that shape a platform's recommendations.

August 2023; Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 Harvard Law Review 524 <https://harvardlawreview.org/wp-content/uploads/2022/11/136-Harv.-L.-Rev.-526.pdf> accessed 22 August 2023; Aziz Z Huq, 'Constitutional Rights in the Machine-Learning State' (2020) 105 Cornell Law Review 1875 <https://chicagounbound.uchicago.edu/journal_articles/10125/> accessed 22 August 2023; Bernhard Rieder, Ariadna Matamoros Fernandez and Oscar Coromina, 'From Ranking Algorithms to "Ranking Cultures": Investigating the Modulation of Visibility in YouTube Search Results' (2018) 24 Convergence 50 <https://eprints.qut.edu.au/223522/> accessed 22 August 2023.

62 DSA, art 39.

63 Leerssen (n 5); Leerssen (n 17); Philip M Napoli, 'Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers' (2015) 39 Telecommunications Policy 751 <https://www.sciencedirect.com/science/article/pii/S030859611400192X> accessed 22 August 2023; Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 Harvard Law Review 524 <https://harvardlawreview.org/wp-content/uploads/2022/11/136-Harv.-L.-Rev.-526.pdf> accessed 22 August 2023; Aziz Z Huq, 'Constitutional Rights in the Machine-Learning State' (2020) 105 Cornell Law Review 1875 <https://chicagounbound.uchicago.edu/journal_articles/10125/> accessed 22 August 2023; Bernhard Rieder, Ariadna Matamoros Fernandez and Oscar Coromina, 'From Ranking Algorithms to "Ranking Cultures": Investigating the Modulation of Visibility in YouTube Search Results' (2018) 24 Convergence 50 <https://eprints.qut.edu.au/223522/> accessed 22 August 2023.

64 DSA 2022, art 14(1).

65 DSA 2022, art 42(2)(a).

66 DSA 2022, art 3(t).

By failing to adequately consider the above aspects, the DSA possibly leaves out more than it might illuminate. A framework that does not account for the socially embedded nature of such systems, and how their interactions with users progressively shapes their functioning (as discussed above), may be particularly inadequate for Global South jurisdictions. As we shall discuss in Chapter V (*Researcher access to platform data*), the lack of access to data held by platforms has impeded empirical research on how platforms influence the propagation of speech and information in such jurisdictions. In the Global North, collaborations between platforms, researchers, and governments have produced rich insights on how recommender systems encode structural discrimination, thereby reproducing racialised discourse (for instance).[67] Conversely, how such systems influence (say) caste-based hierarchies in Global South jurisdictions remains relatively underexplored. In such circumstances, the disclosure of abstract algorithmic parameters is unlikely to contribute to a contextual understanding of the social impacts of recommender systems in such jurisdictions.

b. Looking beyond individual users

While securing more information on recommender systems for individual users may be a step forward, there is scepticism around how far this focus on individual users can assist in holding platforms accountable.

As discussed earlier, the ostensible value of the DSA's user-facing disclosures is premised on the individual user's willingness and capacity to interpret such information and make decisions accordingly. However, this premise is questionable. As observed frequently in the context of privacy-notices, most users do not even read such notices or the terms-and-conditions that provide their basis.[68] Even where users

67 Ariadna Matamaros-Fernandez and Johan Frakas, 'Racism, Hate Speech, and Social Media: A Systematic Review and Critique' (2021) 22(2) Television & New Media 205-224 <https://doi.org/10.1177/1527476420982230> accessed 9 September 2024; Ana-Maria Bliuc and others, 'Online networks of racial hate: A systematic review of 10 years of research on cyber-racism' (2018) 87 Computers in Human Behavior 75-86 <https://doi.org/10.1016/j.chb.2018.05.026> accessed 9 September 2024.

68 Alessandro Mantelero, 'The Future of Consumer Data Protection in the E.U. Rethinking the "Notice and Consent" Paradigm in the New Era of Predictive Analytics' (2014) 30 Computer Law and

read such documents, their complexity and legalistic nature precludes a meaningful understanding of their implications.[69] Thus, scholars apprehend that user-facing disclosures on recommender systems may be similarly constrained in their ability to foster accountability from platforms.[70] On the contrary, they reinforce the burden on individual users to seek and comprehend the disclosures for themselves, while serving as an illusion of greater transparency.[71]

The DSA's individualistic understanding of user-empowerment, as evidenced in its transparency framework for recommender systems, could be particularly ill-suited to the Global South. As discussed earlier, most Global South jurisdictions continue to exhibit low-to-moderate levels of literacy, digital literacy and technical literacy.[72] Thus, many users in such jurisdictions are particularly unlikely to be able to access and derive meaningful insights from disclosures such as those required under the DSA.

Security Review 643 <http://dx.doi.org/10.1016/j.clsr.2014.09.004> accessed 22 August 2023; Joel R Reidenberg and others, 'Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding' (2015) 30 Berkeley Technology Law Journal <https://papers.ssrn.com/abstract=2418297> accessed 22 August 2023.

69 Bernard Schermer and others, 'The Crisis of Consent: How stronger legal protection may lead to weaker consent in data protection' (2014) 16 Ethics and Information Technology 171-182, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2412418>; Mantelero (n 68).

70 Budzinski and Karg (n 7); Jérôme De Cooman, 'Humpty Dumpty and High-Risk AI Systems: The Ratione Materiae Dimension of the Proposal for an EU Artificial Intelligence Act' (2022) VI Market and Competition Law Review <https://orbi.uliege.be/handle/2268/291543> accessed 22 August 2023.

71 Budzinski and Karg (n 7); De Cooman (n 67); Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability (2016) 20 New Media & Society <https://journals.sagepub.com/doi/10.1177/1461444816676645> accessed 22 August 2023; Elettra Bietti, 'Consent as a Free Pass: Platform Power and the Limits of the Informational Turn', 40 (2020) Pace Law Review 307 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489577> accessed 25 January 2025..

72 'With Almost Half of World's Population Still Offline, Digital Divide Risks Becoming "New Face of Inequality", Deputy Secretary-General Warns General Assembly' (*United Nations* 27 April 2021)<https://press.un.org/en/2021/dsgsm1579.doc.htm> accessed 22 August 2023; United Nations Development Programme, 'Submission to the Global Digital Compact' (April 2023) <https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/GDC-submission_UNDP.pdf> accessed 22 August 2023.

c. Looking beyond engagement-optimisation

Notwithstanding their limitations, transparency requirements for algorithmic recommender systems can perhaps contribute to the public understanding of these systems over time. However, it remains to be seen whether the information gathered from such disclosures can cause a broader shift away from the goal of engagement-optimisation.

As discussed earlier, many societal risks arising from platforms have been recognised to stem from (or be compounded by) platforms' commercial drive to optimise for users' attention. As a result of this recognition, public pressure has grown on major platforms to systematically alter their recommending models and orient them towards other goals. Proposals oriented towards incorporating values such as 'diversity',[73] 'agonism'[74] and 'human autonomy'[75] (and their various conceptions and combinations) have been advanced by scholars.[76] In some cases, major platforms have also displayed a commitment towards public values, by entering into voluntary codes of practice and research to explore alternatives to engagement optimisation.[77]

---

73 Laura Schelenz, 'Diversity-Aware Recommendations for Social Justice? Exploring User Diversity and Fairness in Recommender Systems', Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (ACM 2021) <https://dl.acm.org/doi/10.1145/3450614.3463293> accessed 22 August 2023; Matevž Kunaver and Tomaž Požrl, 'Diversity in Recommender Systems – A Survey' (2017) 123 Knowledge-Based Systems 154 <https://linkinghub.elsevier.com/retrieve/pii/S0950705117300680> accessed 22 August 2023; Pablo Castells, Neil Hurley and Saul Vargas, 'Novelty and Diversity in Recommender Systems', Recommender Systems Handbook (Springer 2022) <https://doi.org/10.1007/978-1-0716-2197-4_16> accessed 22 August 2023.

74 Kate Crawford, 'Can an Algorithm Be Agonistic? Ten Scenes from Life in Calculated Publics' (2016) 41 Science, Technology, & Human Values 77 <https://doi.org/10.1177/0162243915589635> accessed 22 August 2023.

75 Lav R Varshney, 'Respect for Human Autonomy in Recommender Systems' (arXiv, 5 September 2020) <http://arxiv.org/abs/2009.02603> accessed 22 August 2023.

76 For a survey of various public values advanced in this context, *see* Stray and others (n 49); *see also* Alvise De Biasio and others, 'A Systematic Review of Value-Aware Recommender Systems' (2023) 226 Expert Systems with Applications 120131 <https://www.sciencedirect.com/science/article/pii/S0957417423006334> accessed 22 August 2023.

77 Global Partnership on AI, 'Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action, Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions' (The Global Partnership on Artificial Intelligence 2022) p 35

More concretely, some platforms have also announced alterations to their recommendation algorithms to diverge from the purely engagement-driven logic, towards value-awareness.[78]

These developments illustrate the deeply political and contentious nature of recommender system governance and more generally, recommending.[79] The design of recommender systems, and thus, their output, is intricately shaped by the economic imperatives of demand and supply, alongside political pressures exerted by regulatory evolution and civil society.[80] The question of what platforms should optimise for (i.e. the nature of content that platforms should promote and demote), has many competing and often conflicting responses.[81] In this context, transparency requirements such as those under the DSA can serve to bring platforms' respective priorities, as reflected in their design choices, into sharper focus. However, in order to push dominant platforms to critically reconsider and eschew their engagement-centred recommending models, challenging their monopolistic powers is critical.

In the *status quo*, even where a well-informed user recognises that a major platform's recommendations are at odds with their values or demands, they are

---

<https://gpai.ai/projects/responsible-ai/transparency-mechanisms-for-social-media-recommender-algorithms.pdf> accessed 22 August 2023.

78 Adam Mosseri [@mosseri], (Instagram) '👇 New Features....' <https://twitter.com/mosseri/status/1516811952235769860> accessed 22 August 2023; Google, 'What Site Owners Should Know about Google's August 2019 Core Update' (Google) <https://developers.google.com/search/blog/2019/08/core-updates> accessed 22 August 2023; Adam Mosseri, 'Bringing People Closer Together' (Facebook, 11 January 2018) <https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/> accessed 22 August 2023; Adam Mosseri, 'Facebook Recently Announced a Major Update to News Feed; Here's What's Changing | Meta' (Facebook, 8 April 2018) <https://about.fb.com/news/2018/04/inside-feed-meaningful-interactions/> accessed 22 August 2023; The YouTube Team, 'Continuing Our Work to Improve Recommendations on YouTube' (*YouTube Official Blog*, 25 January 2019) <https://blog.youtube/news-and-events/continuing-our-work-to-improve/> accessed 22 August 2023.

79Leerssen (n 5).

80 Leerssen (n 5).

81 For instance, a recommender system committed to formal neutrality will throw up a significantly different set of recommendations from a system designed to amplify marginalised voices. Applied to a music streaming platform, the former may recommend albums ranked highly on the Billboard Top100, while the latter may recommend lesser-known independent scores. *See* Stray and others (n 49); Leerssen (n 5).

inhibited from migrating to another platform in view of the inordinate switching-costs.[82] This may be particularly true in many Global South economies, where major platforms exert overwhelming market-dominance, locking-in users and leaving little space for new market-entrants.[83] Thus, a competition framework tailored to digital markets, such as the EU's Digital Markets Act (DMA), can be expected to dilute entry-barriers and foster the emergence of platforms with alternative business models, governance structures and design choices. In parallel, it could facilitate the emergence of third-party algorithms, through which users can explore alternative recommendations.[84] This would provide users in the region a broader range of options, aligned with a broader range of value-systems and socio-political contexts. Concomitantly, Global South states, in attempting to institute user-facing disclosure requirements for recommender systems, must consider other systemic approaches to complement them. Frameworks under the DSA that will be discussed in subsequent chapters of this report – particularly those relating to risk management (*Chapter III*), audits (*Chapter IV*) and researchers' access to platform data (*Chapter V*) – can facilitate a broader, more functional and context-sensitive understanding of recommender systems in such jurisdictions.

82Budzinski and Karg (n 7).

83 Annabelle Gawer and Carla Bonina, 'Digital platforms and development: Risks to competition and their regulatory implications in developing countries' 34(3) (2024) Information and Organization 100525 <https://doi.org/10.1016/j.infoandorg.2024.100525> accessed 25 January 2025; United Nations Conference on Trade and Development, 'Digital Economy Report 2019, Value Creation and Capture: Implications for Developing Countries (United Nations, 2019) <https://unctad.org/publication/digital-economy-report-2019> accessed 25 January 2025.

84Helberger and others (n 36); Maria Luisa Stasi, 'Social media markets: A pro-competitive approach to free speech challenges' (2023) [Doctoral Thesis, Tilburg University] <https://pure.uvt.nl/ws/portalfiles/portal/85262130/Stasi_Social_19-12-2023.pdf> accessed 18 September 2024; Francis Fukuyama and others, 'Middleware for Dominant Digital Platforms: A Technological Solution to a Threat to Democracy' Stanford Cyber Policy Center, 3, <https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/cpc-middleware_ff_v2.pdf> accessed 18 September 2024.

Insights for the Global South

- Users in the Global South exhibit a very limited understanding of the role of recommender systems in delivering content, and many perceive their content-feeds as neutral representations of reality. In this context, disclosure of the use of recommender systems and the parameters used by such systems, can offer users at least an elementary understanding as to why content appears to them in the given manner and priority.

- Nonetheless, parametric disclosures (as required by the DSA) cannot, by themselves, explain how recommender systems interact with the information ecosystems in which they operate. Such systems are fundamentally socio-technical in nature, and their outputs are shaped by platforms' design-choices, organisational processes as well as users' behaviour, alongside other key factors. Other (more systemic) transparency mechanisms, such as audits, risk assessments and researcher access to data, are expected to be instrumental in shedding more light on such factors.

- The DSA's user-facing disclosures assume an empowered user, capable of and willing to interpret such disclosures and make decisions accordingly. However, given the low-to-moderate levels of literacy, digital literacy and technical literacy currently prevailing in Global South jurisdictions, many users are particularly unlikely to be able to access and derive meaningful insights from such disclosures.

- Major platforms' advertising-based business models, based on optimising for users' engagement, are central to many risks that they create or propagate. In recent years, political pressure has forced certain platforms to incorporate other considerations, such as diversity and factual accuracy, in recommending content. Transparency requirements can bring platforms' priorities and choices into sharper focus. However, they must be complemented with measures in competition law to secure for users the effective choice to move to

other platforms, if the content recommended by a platform does not align with their priorities and value-systems.

- The DSA requires VLOPs and VLOSEs to provide at least one option to users to access recommended content without profiling them. However, such optionality may not be very meaningful, particularly in the Global South, where most users may only have a basic understanding of how recommender systems operate and of the broader risks posed by profiling. Thus, platforms should not be permitted to profile a user to deliver recommended content, until and unless the user has expressly and affirmatively consented to it, after being informed of the associated risks in adequate detail, in a manner that is understandable and clearly accessible.

- The collection and use of personal data lie at the heart of personalised recommendations. Pertinently, DP laws provide controls and place safeguards on the collection and use of personal data by any entity, including platforms. While DP laws are being increasingly adopted since the introduction of the GDPR, many Global South jurisdictions are yet to enact such a law. Thus, Global South jurisdictions contemplating transparency frameworks for recommender systems must, on priority, adopt DP laws to undergird them.

2. TRANSPARENCY IN ADVERTISING

2.1. Introduction

Advertising constitutes an essential component of platform economics,[1] and for the longest time, it has operated with almost no transparency and little oversight.[2] The 2016 United States Presidential Election was a watershed moment for platform transparency, [3] leading to heightened public and regulatory scrutiny around advertisements.[4] As evidence of a concerted disinformation campaign by Russia's Internet Research Agency (IRA) mounted, the largest social media platforms faced intense public backlash.[5] The IRA had extensively mobilised paid advertisements on Facebook and Instagram, trying to target and manipulate voters based on their ethnicity and political leanings.[6] In response to the resulting backlash, platforms

1 Sarah Myers West, 'Data Capitalism: Redefining the Logics of Surveillance and Privacy' (2019) 58 Business & Society 20 <https://doi.org/10.1177/0007650317718185> accessed 7 February 2024; See also James Ball, 'Online Ads Are About to Get Even Worse' *The Atlantic* (1 June 2023) <https://www.theatlantic.com/technology/archive/2023/06/advertising-revenue-google-meta-amazon-apple-microsoft/674258/> accessed 8 February 2024.

2 José Estrada-Jiménez and others, 'Online Advertising: Analysis of Privacy Threats and Protection Approaches' (2017) 100 Computer Communications 32 <https://linkinghub.elsevier.com/retrieve/pii/S0140366416307083> accessed 7 February 2024.

3 Robert Gorwa and Timothy Garton Ash, 'Democratic Transparency in the Platform Society', *Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge University Press 2020) <https://www.cambridge.org/core/books/social-media-and-democracy/democratic-transparency-in-the-platform-society/F4BC23D2109293FB4A8A6196F66D3E41> accessed 10 November 2023.

4 Márcio Silva and others, 'Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook' (arXiv, 31 January 2020) <http://arxiv.org/abs/2001.10581> accessed 12 May 2023.

5 ibid.

6 These polarising and divisive advertisements targeted different groups differently. They encouraged conservative voters to vote for Trump while voters from the African American communities were encouraged to boycott the elections. See Philip N Howard and others, 'The IRA, Social Media and Political Polarization in the United States, 2012-2018' <https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2018/12/The-IRA-Social-Media-and-Political-Polarization.pdf> accessed 11 December 2023.

adopted several voluntary transparency measures, including disclosures about online advertisements, to regain public trust.[7] Facebook introduced a public library for political advertisements in the US,[8] followed by Google[9] and Twitter.[10] Other voluntary transparency measures, like identity verification for political advertisements, followed suit.[11] Notably, these mechanisms were first introduced in select countries before being rolled out globally.[12] Regulatory initiatives like the US Honest Ads Bill,[13] Canada's Elections Modernization Act,[14] and Ireland's Online Advertising and Social Media (Transparency) Bill[15] had also emerged in this context.[16]

---

7 Gorwa and Ash (n 3).

8 David Ingram, 'Facebook Launches Searchable Archive of U.S. Political Ads' *Reuters* (24 May 2018) <https://www.reuters.com/article/idUSKCN1IP37H/> accessed 8 February 2024.

9 Taylor Hatmaker, 'Google Releases a Searchable Database of US Political Ads' *TechCrunch* (15 August 2018) <https://techcrunch.com/2018/08/15/google-political-ad-library/> accessed 8 February 2024.

10 Bruce Falck, 'Providing More Transparency around Advertising on Twitter' (*X Blog*, 28 June 2018) <https://blog.twitter.com/en_us/topics/company/2018/Providing-More-Transparency-Around-Advertising-on-Twitter> accessed 8 February 2024.

11 See for instance, Meta, 'Bringing More Transparency to Political Ads in 2019' (*Meta for Business*) <https://www.facebook.com/business/news/bringing-more-transparency-to-political-ads-in-2019> accessed 28 April 2023; 'Availability for Ads about Social Issues, Elections or Politics' (*Meta Business Help Centre*) <https://en-gb.facebook.com/business/help/215015729527 6323> accessed 8 February 2024; 'Increasing Transparency through Advertiser Identity Verification' (*Google Ads & Commerce Blog*, 23 April 2020) <https://blog.google/products/ads/advertiser-identity-verification-for-transparency/> accessed 8 February 2024.

12 See for instance, Paresh Dave, 'Exclusive - Facebook Brings Stricter Ads Rules to Countries with Big 2019 Votes' *Reuters* (16 January 2019) <https://www.reuters.com/article/idUSKCN1PA0C5/> accessed 8 February 2024.

13 Byron Tau, 'Proposed "Honest Ads Act" Seeks More Disclosure About Online Political Ads' *Wall Street Journal* (19 October 2017) <https://www.wsj.com/articles/proposed-honest-ads-act-seeks-more-disclosure-about-online-political-ads-1508440260> accessed 27 May 2024.

14 Michael Pal, 'Evaluating Bill C-76: The Elections Modernization Act' (25 August 2019) <https://papers.ssrn.com/abstract=3572737> accessed 27 May 2024.

15 See Online Advertising and Social Media (Transparency) Bill 2017 <https://www.oireachtas.ie/en/bills/bill/2017/150/> (now lapsed).

16 Paddy Leerssen and others, 'Platform Ad Archives: Promises and Pitfalls' (2019) 8 Internet Policy Review <https://policyreview.info/articles/analysis/platform-ad-archives-promises-and-pitfalls> accessed 4 May 2023.

Historically, both commercial and political advertising in the media have been regulated.[17] However, online advertising presents unique challenges to regulation, including addressing harms arising from personalisation and micro-targeting.[18] These are compounded by the complexity and opacity of the online advertising ecosystem,[19] and the dominance of a few key players.[20] Even monitoring and enforcement of regulation is challenging given the speed and scale of online communication, the comparative ease of cross-border advertising and the use of proxies to buy ads.[21]

Advertisers infer and use insights from user information gathered from various sources, including their behaviour on platforms, to profile and segment users for commercial targeting.[22] This can contain personally identifiable information, including sensitive information such as political preferences, health-related data, and other demographic markers.[23] Such data can potentially be used by malicious advertisers to discriminate based on ethnicity, gender, sexuality, and political preferences.[24] Moreover, algorithmic bias and the economic logic of automated

17 Many jurisdictions have disclosure and reporting requirements for election advertisements, as well as campaigning caps and silence periods prior to elections. Commercial advertisements are also subject to consumer protection, copyright, and competition laws. See ibid.

18 Athanasios Andreou and others, 'Measuring the Facebook Advertising Ecosystem', *Proceedings 2019 Network and Distributed System Security Symposium* (Internet Society 2019) <https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_04B-1_Andreou_paper.pdf> accessed 2 January 2024; Sara Kingsley and others, 'Auditing Digital Platforms for Discrimination in Economic Opportunity Advertising'.

19 Estrada-Jiménez and others (n 2).

20 See Patience Haggin, 'Google and Meta's Advertising Dominance Fades as TikTok, Netflix Emerge' *mint* (3 January 2023) <https://www.livemint.com/industry/advertising/google-and-meta-s-advertising-dominance-fades-as-tiktok-netflix-emerge-11672749572663.html> accessed 7 February 2024.

21 Leerssen and others (n 16).

22 Estrada-Jiménez and others (n 2); West (n 1); Wes Davis, 'This Is How Facebook Knows Where You've Been and What You Bought' *The Verge* (17 January 2024) <https://www.theverge.com/2024/1/17/24041897/facebook-meta-targeted-advertising-data-mining-study-privacy> accessed 7 February 2024.

23 Estrada-Jiménez and others (n 2).

auctioning systems can perpetuate historical inequalities in displaying advertisements for job opportunities, credit, housing, etc.[25]

The prevalence and impact of such systems could be even more acute in countries that lack adequate anti-discrimination laws in employment and housing. Further, in the absence of effective data protection frameworks prohibiting profiling based on sensitive categories like ethnicity, gender, sexuality, and religion, advertisers can engage in unchecked and invasive targeting.

The most egregious harms of such systemic discrimination can be witnessed in political microtargeting, where distortion and fragmentation of the public discourse can be mobilised for voter manipulation, polarisation, and spreading disinformation.[26]

It is in this context that the EU's Digital Services Act (DSA) lays down obligations for advertisement transparency. It mandates Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to maintain public advertisement repositories (accessible through a user interface tool as well as API).[27] All online platforms[28] (including VLOPs and VLOSEs) are obligated to provide: (a) user-facing advertisement disclaimers (including information on sponsors and targeting

---

24 Andreou and others (n 18); Till Speicher and others, 'Potential for Discrimination in Online Targeted Advertising', *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (PMLR 2018) <https://proceedings.mlr.press/v81/speicher18a.html> accessed 1 February 2024; Julia Angwin Tobin Madeleine Varner, Ariana, 'Facebook Enabled Advertisers to Reach "Jew Haters"' (*ProPublica*, 14 September 2017) <https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters> accessed 7 February 2024.

25 See Anja Lambrecht and Catherine Tucker, 'Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads' (2019) 65 Management Science 2966 <https://pubsonline.informs.org/doi/10.1287/mnsc.2018.3093> accessed 5 February 2024; Kingsley and others (n 18); Amit Datta, Michael Carl Tschantz and Anupam Datta, 'Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination' (arXiv, 16 March 2015) <http://arxiv.org/abs/1408.6491> accessed 7 February 2024.

26 Frederik Borgesius and others, 'Online Political Microtargeting: Promises and Threats for Democracy' (2018) 14 Utrecht Law Review <https://utrechtlawreview.org/articles/10.18352/ulr.420>.

27 DSA 2022, arts 39(1), (2) and (3).

28 All online platforms except MSMEs that are not VLOPs/VLOSEs.

attributes);[29] (b) user-control of advertisement targeting (through option to change targeting parameters); [30] and (c) public disclosure for user-generated paid promotion/influencer marketing.[31]

2.2. Public Advertisement Repositories

The personalisation of advertisements results in structural informational asymmetry, as only those targeted by a particular ad are exposed to it.[32] As a result, there exists no method for systemic scrutiny or monitoring of the overall advertisement ecosystem. This makes public advertisement repositories an important step towards achieving greater transparency.

As discussed, post-2016, many platforms voluntarily made available advertisement repositories, serving as historical databases of ads displayed on their platforms. These typically included information on the advertiser, the cost incurred on the ad and some statistics on estimated reach.[33] These repositories allow a diversity of stakeholders, including regulators, journalists, civil society organisations, rival advertisers and users, to monitor the platform advertising ecosystem.[34] They have used these voluntary repositories to hold politicians accountable, investigate corporate and political astroturfing,[35] uncover violations of campaign funding laws, fact-check campaign ads, and monitor hate speech.[36]

---

29 DSA 2022, art 26(1).

30 DSA 2022, art 26(1)(iv).

31 DSA 2022, art 26(2).

32 Paddy Leerssen, 'Algorithm Centrism in the DSA's Regulation of Recommender Systems' [2022] Verfassungsblog<https://verfassungsblog.de/roa-algorithm-centrism-in-the-dsa/> accessed 1 May 2023.

33 See for instance Meta's voluntary ad library. 'Ad Library' (*Meta*) <https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=IN&sort_data[direction]=desc&sort_data[mode]=relevancy_monthly_grouped&media_type=all> accessed 19 May 2024.

34 Leerssen and others (n 16).

35 Kovic et al. define digital astroturfing as "a form of manufactured, deceptive and strategic top-down activity on the Internet initiated by political actors that mimics bottom-up activity by autonomous individuals". Marko Kovic and others, 'Digital Astroturfing in Politics: Definition,

Though these voluntary archives have proven to be a welcome initiative for advertisement transparency, they suffer from issues of quality and reliability and are marred by technical glitches, loss of historical data,[37] and problems of over-inclusion and under-inclusion.[38] They also typically do not contain information on targeting by advertisers, limiting the scope of accountability derived from them.[39]

Thus, mandating and regulating advertisement repositories by legislation may potentially result in more reliable databases, including additional information that platforms typically do not disclose in their voluntary repositories.

The DSA mandates VLOPs and VLOSEs to make publicly available a repository containing all advertisements being presented to users at any given time, along with ads presented over the previous year, through a "searchable and reliable tool that allows multi-criteria queries and through application programming interfaces." and to "make reasonable efforts to ensure that the information is accurate and complete."[40] These repositories shall include information on the content or the subject matter of the advertisement,[41] the person on whose behalf the advertisement is presented,[42] and the person who has paid for the ad,[43] the period for which the ad

---

Typology, and Countermeasures' (2018) 18 Studies in Communication Sciences 69 <https://www.researchgate.net/publication/332867727_Digital_astroturfing_in_politics_Definition_typology_and_countermeasures> accessed 23 November 2023.

36 Paddy Leerssen and others, 'News from the Ad Archive: How Journalists Use the Facebook Ad Library to Hold Online Advertising Accountable' (2021) 0 Information, Communication & Society 1 <https://doi.org/10.1080/1369118X.2021.2009002> accessed 28 April 2023.

37 Cynthia O'Murchu, Jemima Kelly and David Blood, 'Facebook under Fire as Political Ads Vanish from Archive' *Financial Times* (10 December 2019) <https://www.ft.com/content/e6fb805e-1b78-11ea-97df-cc63de1d73f4> accessed 3 May 2023.

38 'Facebook's Ad Archive API Is Inadequate | The Mozilla Blog' <https://blog.mozilla.org/en/mozilla/facebooks-ad-archive-api-is-inadequate/> accessed 3 May 2023.

39 Leerssen and others (n 36).

40 DSA 2022, art 39(1).

41 DSA 2022, art 39(2)(a).

42 DSA 2022, art 39(2)(b).

43 DSA 2022, art 39(2)(c).

was being presented on the platform,[44] whether the ad was targeted to particular groups and the main parameters used for such targeting,[45] the total number of recipients reached and where applicable disaggregated data broken down by member states for targeted groups.[46]

This regulation through the DSA is likely to keep a check on the quality and accuracy of archives while also protecting them from changing corporate policies.[47] The DSA mandates platforms to "make reasonable efforts to ensure that the information is accurate and complete".[48] However, it is worth noting that the DSA-mandated repository user interface and APIs will be provided and controlled by the platforms themselves.[49] The challenges of regulating these repositories to ensure meaningful transparency have already come to the fore. A recent study of major platforms conducted by Mozilla and Check First assessed the readiness and compliance of ad repositories against Article 39 of the DSA and good practice guidelines authored by independent experts.[50] The study highlights substantial deficiencies in the accessibility and functionality of the APIs and user interface of repositories, completeness and granularity of data, and documentation for research.[51]

Improving the quality and reliability of the repositories compared to their voluntary predecessors would entail a combination of regulatory oversight, periodic monitoring for compliance, reporting mechanisms for users and researchers, and, most

44 DSA 2022, art 39(2)(d).

45 DSA 2022, art 39(2)(e).

46 DSA 2022, art 39(2)(g).

47 See for instance, Jessica Piper, 'Twitter Fails to Report Some Political Ads after Promising Transparency' (*POLITICO*, 10 April 2023) <https://www.politico.com/news/2023/04/10/twitter-political-ads-transparency-00091077> accessed 13 May 2023.

48 DSA 2022, art 39(1).

49 Mozilla EU Policy, 'Mozilla Position Paper on the EU Digital Services Act' <https://blog.mozilla.org/netpolicy/files/2021/05/Mozilla-DSA-position-paper-.pdf> accessed 9 May 2023.

50 Mozilla and Check First, 'Full Disclosure: Stress Testing Tech Platforms' Ad Repositories' (2024) <https://assets.mofoprod.net/network/documents/Full_Disclosure_Stress_Testing_Tech_Platforms_Ad_Repositories_3FepU2u.pdf> accessed 14 August 2024.

51 ibid.

importantly, platform cooperation. Here, the state's regulatory capacity, as well as, platforms' willingness to invest resources, becomes paramount. This may prove to be especially challenging in many Global South countries, which are not priority markets for platforms.[52] While mandating archives can be beneficial to Global South countries, strict regulation may be difficult to implement in practice given the lobbying and pushback that often accompany such measures.[53] Big Tech companies have, on various occasions, threatened to leave the market or block parts of their service when confronted with regulations they deem unacceptable.[54]

In terms of scope, the ad repositories mandated by the DSA go further than most existing voluntary libraries by not being limited to political advertisements. This not only overcomes the challenges of defining and identifying political ads at scale but

52 For instance, Facebook allocated content moderation resources to countries based on a hierarchical tier-based system which left a majority of the world with little oversight. See Ben Gilbert, 'Facebook Ranks Countries into Tiers of Importance for Content Moderation, with Some Nations Getting Little to No Direct Oversight, Report Says' *Business Insider* (5 October 2021) <https://www.businessinsider.in/tech/news/facebook-ranks-countries-into-tiers-of-importance-for-content-moderation-with-some-nations-getting-little-to-no-direct-oversight-report-says/articleshow/87263447.cms> accessed 17 May 2023.

53 It has been reported that Facebook lobbied against the Honest Ads Act and preemptively implemented voluntary transparency mechanisms to forestall regulation in the US. See Heather Timmons Kozlowska Hanna, 'Facebook's Quiet Battle to Kill the First Transparency Law for Online Political Ads' (*Quartz*, 22 March 2018) <https://qz.com/1235363/mark-zuckerberg-and-facebooks-battle-to-kill-the-honest-ads-act> accessed 9 May 2023; Facebook also lobbied against strict rules on online advertisements during Indian elections. See Deeksha Bhardwaj and Venkat Ananth, 'Facebook Lobbied over Poll Rules: Papers' *Hindustan Times* (New Delhi, 22 November 2021) <https://www.hindustantimes.com/india-news/fb-lobbied-over-poll-rules-papers-101637530999072.html>.

54 Major Global online platforms, including, Facebook, Google and Twitter, threatened to exit Pakistan as it proposed stringent data localisation and content takedown laws. However, Singh notes, similar proposals in neighbouring India were not met with equivalent backlash by Big Tech which considers India to be an important market. See Manish Singh, 'Google, Facebook and Twitter Threaten to Leave Pakistan over Censorship Law' (*TechCrunch*, 20 November 2020) <https://techcrunch.com/2020/11/20/google-facebook-and-twitter-threaten-to-leave-pakistan-over-censorship-law/> accessed 28 September 2022; However, this trend is not limited to Global South. Meta started blocking news in Canada in response to a law that mandated compensating news organizations. See Katie Robertson, 'Meta Begins Blocking News in Canada' *The New York Times* (2 August 2023) <https://www.nytimes.com/2023/08/02/business/media/meta-news-in-canada.html> accessed 6 February 2024.

also provides much-needed transparency on commercial advertising.[55] The ad repositories under the DSA also include user-generated ads or influencer ads, which are emerging as an important means for brands and political actors to reach a larger audience.[56] The repositories must include metadata on ads that platforms have removed or disabled access to on grounds of illegality or violation of their Terms and Conditions.[57] Metadata on ads that have been removed/blocked either by the platform's own voluntary action or on receiving notice, must include the statements of reasons explaining the legal or contractual ground allegedly violated and the facts and circumstances relied on in taking the decision. For ads that have been removed/blocked pursuant to a state order, the repository must contain information on the legal basis as outlined in the order. These can be a good first step towards understanding how platforms monitor and moderate ads.[58]

As noted previously, the DSA lays down the metadata to be included in the archive, notably including information on targeting parameters.[59] Through this provision, the DSA takes a giant leap forward in mandating disclosure of the main parameters for targeting that were generally missing from voluntary repositories by platforms. However, scholars believe that the language in the DSA still leaves some scope for platforms to withhold vital information on targeting through their interpretation of what constitutes "main parameters"[60] (also see Chapter 1 for a detailed discussion on

55 Leerssen and others (n 16); Aaron Rieke and Miranda Bogen, 'Leveling the Platform: Real Transparency for Paid Messages on Facebook'.

56 DSA 2022, art 39(2)(f).

57 DSA 2022, art 39(3).

58 Additionally, Leerssen suggests that additional information including buyer identity and ad content could be made available for ads that are taken down for violating the platform's TOS but are not illegal. This can go a long way in understanding how platform content moderation operates. See Paddy Leerssen, 'Platform Ad Archives in Article 30 DSA' (*DSA Observatory*, 25 May 2021) <https://dsa-observatory.eu/2021/05/25/platform-ad-archives-in-article-30-dsa/> accessed 10 May 2023.

59 Article 39(2) lays down information to be included in the archives: (a) content of the ad; (b) sponsor and buyer information; (c) time period for which the ad was active; (d) main parameters used for targeting including any exclusion criteria; (e) influencer ad details; (f) reach data segregated by targeted categories and member states

60 Paddy Leerssen (n 58).

disclosing "main parameters" for recommending). Such concerns become more prominent in the face of research that demonstrates Facebook's voluntary "ad preferences" explanations being replete with incomplete and misleading data.[61]

A more beneficial way of seeking targeting information would be to ensure that archives have an equivalent level of targeting information in terms of categories and granularity as is available to advertisers.[62] Additionally, information on a/b testing,[63] whether targeting data was based on platform-defined user interests, or advertiser custom lists[64] or characteristics sourced from data brokers[65] can be useful for researchers, civil society actors and regulators across the globe. This is especially significant because advertisers can potentially discriminate against users based on ethnicity, race, gender and other sensitive parameters even without explicitly using these discriminatory attributes for targeting.[66] Research suggests that using features like custom lists where personally identifiable information (PII) (such as phone numbers or email addresses) is directly entered by the advertiser or constructing look-alike audiences[67] on platforms like Meta can lead to discriminatory advertising

61 Athanasios Andreou and others, 'Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations' (2018).

62 Paddy Leerssen (n 58); Mozilla, 'Facebook and Google: This Is What an Effective Ad Archive API Looks Like' (*The Mozilla Blog*, 28 March 2019) <https://blog.mozilla.org/en/mozilla/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/>.

63 A/B testing consists of testing a hypothesis with a control (version A) and a treatment (version B). In advertising, these versions can consist of different variables like targeting parameters, advertisement content etc and the experiment can be used to test the cost-effectiveness or reach of different advertising strategies. See Ron Kohavi and Roger Longbotham, 'Online Controlled Experiments and A/B Tests' in Dinh Phung, Geoffrey I Webb and Claude Sammut (eds), *Encyclopedia of Machine Learning and Data Science* (Springer US 2023) <https://link.springer.com/10.1007/978-1-4899-7502-7_891-2> accessed 27 May 2024; 'About A/B Testing' (*Meta Business Help Centre*) <https://en-gb.facebook.com/business/help/1738164643098669> accessed 27 May 2024.

64 Mozilla (n 62).

65 Julia Angwin, Surya Mattu, and Terry Parris Jr, 'Facebook Doesn't Tell Users Everything It Really Knows About Them' (*ProPublica*, 27 December 2016) <https://www.propublica.org/article/facebook-doesnt-tell-users-everything-it-really-knows-about-them> accessed 13 May 2023.

66 Speicher and others (n 24).

67 Lookalike audience is a feature on Facebook that "leverages information such as demographics, interests and behaviours from your source audience to find new people who share similar qualities". See 'About Lookalike Audiences' (Meta) <https://en-gb.facebook.com/business/help/164749007013531?id=401668390442328>.

even where the use of sensitive attributes for targeting is prohibited by regulation.[68] Using publicly available voter data in the US as input information for Facebook's custom audience, researchers were able to create highly targeted advertisements skewed towards specific races and genders.[69] Researchers have also found that a significant portion of advertisements on Facebook are targeted using features which enable advertisers to input personally identifiable information.[70] Thus, disclosing information beyond "main targeting parameters" is essential for understanding how ad targeting may result in discriminatory outcomes for users.

Thus, while the DSA takes a step forward and mandates information disclosure on advertisement targeting, it has its shortcomings. What is notably missing from the advertisement repositories mandated under the DSA is financial information on the quantum of spending. This appears to be a step back,[71] given most voluntary archives already share this information, and journalists across the world have used this to hold platforms and advertisers accountable.[72] Similarly, there seems to be no rationality for data retention being limited to merely one year,[73] making historical research or regulatory investigations into older ads untenable.[74]

Further, the ambiguities associated with terms like "main parameters" for targeting[75] can be potentially exaggerated in implementation across Global South countries where power dynamics between states and platforms play out differently. Platforms are not incentivised to allocate their resources to many Global South countries and

68 Speicher and others (n 24).

69 ibid.

70 Andreou and others (n 18).

71 Paddy Leerssen (n 58).

72 See for instance, Nayantara Ranganathan and Kumar Sambhav, 'Facebook Charged BJP Less for India Election Ads than Others' *Al Jazeera* (New Delhi, India, 16 March 2022) <https://www.aljazeera.com/economy/2022/3/16/facebook-charged-bjp-lower-rates-for-india-polls-ads-than-others> accessed 12 May 2023.

73 DSA 2022, art 39(1). By contrast, Facebook's voluntary ad libraries provide information on social issues, elections or politics for the past seven years. 'Ad Library' (n 33)

74 Paddy Leerssen (n 58).

75 DSA 2022, art 39(2)(e).

may not be willing to maintain detailed ad repositories.[76] States, too, have limited regulatory capacity to monitor and audit the adequacy of the information, including targeting parameters, disclosed in these repositories. Furthermore, it can be more difficult for citizens and civil society in the Global South to bring proceedings against them in local courts, given that platforms often claim that these courts do not have jurisdiction over them.[77]

Mandating ad repositories for VLOPs and VLOSEs can be an important step in holding platforms accountable in the Global South. However, it is important that the information included in these repositories — such as the methods employed to target users (e.g. custom lists or targeting attributes) and the monetary spending on an advertisement— be carefully deliberated upon and laid out in laws or delegated acts. These must be decided through multi-stakeholder discussions where the voices of civil society, citizens, and researchers, must be adequately represented. Such deliberations can also pave the way for developing standards for ad repositories across platforms. Mozilla's recent study on "stress testing" ad repositories also recommends developing guidelines to establish minimum standards and ensure some degree of harmonisation to ease cross-platform research.[78] Standards and guidelines should leave enough flexibility to accommodate the diversity of platforms while at the same time ensuring that a minimum level of meaningful information disclosure and operational reliability is maintained across platforms. It is also important to note that maintaining repositories can be resource-intensive and imposing such an obligation on smaller platforms might create barriers to entry in a market that is already dominated by a few players.

76 See Zahra Takhshid, 'Regulating Social Media in the Global South' 24; Billy Perrigo, 'Meta Sued Over Ethnic Violence in Ethiopia' [2022] *TIME*<https://time.com/6240993/facebook-meta-ethiopia-lawsuit/> accessed 7 February 2024; Ben Gilbert (n 52); 'YouTube Approves Disinformation Ads in India Ahead of General Election' (*Access Now*) <https://www.accessnow.org/press-release/youtube-disinformation-ads-india-election-2024-en/> accessed 27 May 2024.

77 Takhshid (n 76); Annie Njanja, 'Meta Wants Lawsuit in Kenya Dropped' (*TechCrunch*, 9 June 2022) <https://techcrunch.com/2022/06/09/meta-says-kenyan-court-has-no-jurisdiction-to-determine-case-against-it-wants-it-thrown-out/> accessed 7 February 2024.

78 Mozilla and Check First (n 50).

Finally, accountability derived from archives is highly dependent on a critical and empowered audience and the presence of civil society watchdogs and journalists who can investigate content and flag inconsistencies and illegalities for lawmakers or the general public.[79] Though there exist instances of journalists using the voluntary archives by platforms to hold them accountable in the Global South,[80] these experiences might not be uniform across all countries. Further, it is imperative to consider the structural power asymmetry between civil society and Big Tech, especially in the Global South. This is significant given that accountability derived from journalistic investigations on ad repositories relies on powerful stakeholders like the platforms, advertisers and regulators acknowledging irregularities and taking corrective steps. [81]

2.3. User-Facing Disclaimers

At the heart of online advertising is information asymmetry, with advertisers using insights from personal data to target users, while users remain unaware of the processes behind such ad targeting.[82] This entire process of advertising, from user profiling and classification to the advertiser's choice of targeting methods to the process of bidding, happens in the background without any oversight and accountability to the users.[83] This makes it critical to examine transparency measures that disclose information to users regarding the ads displayed to them. Article 26 of the DSA mandates online platforms to provide user-facing disclaimers to facilitate

79 ibid; Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973 <https://doi.org/10.1177/1461444816676645> accessed 28 February 2023.

80 See Nayantara Ranganathan and Kumar Sambhav, 'How a Reliance-Funded Firm Boosts BJP's Campaigns on Facebook' Al Jazeera (New Delhi, India, 14 March 2022) <https://www.aljazeera.com/economy/2022/3/14/how-a-reliance-funded-company-boosts-bjps-campaigns-on-facebook>.

81 Leerssen and others (n 36).

82 Tom Dobber and others, 'Shielding Citizens? Understanding the Impact of Political Advertisement Transparency Information' [2023] New Media & Society 14614448231157640 <https://doi.org/10.1177/14614448231157640> accessed 12 May 2023.

83 Estrada-Jiménez and others (n 2); West (n 1).

user transparency and user control by identifying sponsored content as advertisements. These disclaimers must include sponsor information, including the entity that bought the ad and paid for the ad, as well as, the "main parameters" used for targeting.[84]

User-facing disclaimers are especially relevant given digital advertising for the past decade has moved towards native advertising, with advertisers preferring inconspicuous ads seamlessly integrated into users' content feeds.[85] More recently, "authentic" advertising in the form of user-generated paid promotion or influencer ads has gained currency.[86]

With respect to such user-generated paid content or influencer advertising, the DSA has mandated platforms to provide users with the functionality to disclose commercial communications and corresponding identifiers for audiences to recognise such sponsored content.[87] Through this provision, the DSA takes an important first step in mandating transparency for influencer content. However, the definition of "commercial communication"[88] for user disclosure and archival does not appear to cover monetised political content by influencers,[89] which is fast

84 Article 26(1) lays down the obligation to ensure that users are able to identify clearly, concisely, unambiguously and in real-time: (i) that the information is an advertisement, including through prominent markings; (ii) the natural or legal person on whose behalf the information is presented; (iii) the natural of legal person who paid for the advertisement; and (iv) meaningful information directly and easily accessible about the main parameters used to determine the users to whom the advertisement is presented and where applicable, how to change those parameters.

85 Cooper Smith, 'The Rise Of Native: Why Social Media Advertising Is Going In-Stream' (*Business Insider*) <https://www.businessinsider.com/the-rise-of-native-advertising-2013-10> accessed 16 May 2023.

86 Kat Shee, 'The Rise Of Influencers In Media' [2023] *Forbes*<https://www.forbes.com/sites/theyec/2023/06/23/the-rise-of-influencers-in-media/>.

87 DSA 2022, art 26(2).

88 Here "commercial communication" refers to "any form of communication designed to promote, directly or indirectly, the goods, services or image of a company, organisation or person pursuing a commercial, industrial or craft activity or exercising a regulated profession." as defined in the Electronic Commerce Directive (Directive 2000/31/EC).

89 Catalina Goanta, 'Human Ads Beyond Targeted Advertising: Content monetization as the blind spot of the Digital Services Act' [2021] Verfassungsblog<https://verfassungsblog.de/power-dsa-dma-11/> accessed 14 May 2023.

emerging as an important mechanism deployed in political campaigns,[90] even in the Global South.[91] The monetisation of political speech on social media presents new challenges for regulation as it often blurs the boundaries between political and commercial speech.[92] Providing information disclosures in the form of disclaimers and archiving such advertisements in repositories could be useful steps to providing transparency. However, it can be challenging to distinguish political speech based on personal convictions from that based on commercial agreements, and thus, enforcing disclosure can also be challenging for both platforms and regulators.[93]

Effectiveness of User-Facing Disclaimers

Mandating labels that help distinguish sponsored content from regular content can be useful, as advertisements are often difficult to spot,[94] especially for first-time

90 Stephanie Lai, 'Campaigns Pay Influencers to Carry Their Messages, Skirting Political Ad Rules' *The New York Times* (2 November 2022) <https://www.nytimes.com/2022/11/02/us/elections/influencers-political-ads-tiktok-instagram.html> accessed 16 May 2023; Magdalena Riedl and others, 'The Rise of Political Influencers—Perspectives on a Trend Towards Meaningful Content' (2021) 6 Frontiers in Communication <https://www.frontiersin.org/articles/10.3389/fcomm.2021.752656> accessed 16 May 2023.

91 Srishti Jaswal, 'Indian Politicians Embrace Influencers Ahead of 2024 Elections' (*Rest of World*, 24 July 2023) <https://restofworld.org/2023/india-2024-elections-influencers/> accessed 8 September 2024; PTI, 'Pakistan Polls: Social Media Playing a Big Role in the Run-up to Feb 8 Polling Day' (*The Print*, 7 February 2024) <https://theprint.in/world/pakistan-polls-social-media-playing-a-big-role-in-the-run-up-to-feb-8-polling-day/1957873/> accessed 7 February 2024; 'BJP Bets on 50 Social Media Influencers for an Edge Online' (*The Indian Express*, 29 November 2022) <https://indianexpress.com/article/cities/delhi/bjp-bets-on-50-social-media-influencers-for-an-edge-online-8294992/> accessed 16 May 2023; Lai (n 90); Chiagozie Nwonwu, Fauziyya Tukur, and Yemisi Oyedepo, 'Nigeria Elections 2023: How Influencers Are Secretly Paid by Political Parties' *BBC News* (18 January 2023) <https://www.bbc.com/news/world-africa-63719505> accessed 7 February 2024; 'Latin American Politicians Court Social-Media Stars, Often Ineptly' *The Economist*<https://www.economist.com/the-americas/2022/07/21/latin-american-politicians-court-social-media-stars-often-ineptly> accessed 7 February 2024.

92 Giovanni De Gregorio and Catalina Goanta, 'The Influencer Republic: Monetizing Political Speech on Social Media' [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3725188> accessed 27 September 2024.

93 ibid.

94 Irina Dykhne, 'PERSUASIVE OR DECEPTIVE? NATIVE ADVERTISING IN POLITICAL CAMPAIGNS' 91.

internet users. User-facing disclaimers aim to increase and activate the persuasion knowledge of users to defend their interests and shield themselves against manipulation or deception.[95] The disclaimers, as per DSA, should also include information on the sponsor, which is particularly useful to prevent corporate astroturfing and manipulation of voters in the case of political and issue-based advertisements. In general, citizens have found information on political sponsors to be empowering as it enables them to gauge the motivation of the campaigner and the lobby groups backing the candidate.[96]

Another important provision is the disclosure of the targeting parameters with the advertisement disclaimer. Besides increasing persuasion knowledge, this can enable users to become aware of the privacy violations arising from targeting. For the motivated or curious user, the DSA also provides the opportunity to manage the parameters for advertisement targeting by providing "meaningful information, where applicable, about how to change" the main parameters for targeting.[97]

However, the effectiveness of user-facing disclaimers is contested at best. Often, the labels or disclaimers go unnoticed by users,[98] and studies show how labelling alone might not be sufficient to help users distinguish sponsored content[99] as mediating factors like digital literacy play an important role[100] in activating the persuasion knowledge of users. In order to be effective, user-facing disclaimers should be

95 Dobber and others (n 82).

96 Dykhne (n 94).

97 DSA 2022, art 26(1)(iv).

98 Sophie C Boerman, Lotte M Willemsen and Eva P Van Der Aa, '"This Post Is Sponsored" Effects of Sponsorship Disclosure on Persuasion Knowledge and Electronic Word of Mouth in the Context of Facebook' (2017) 38 Journal of Interactive Marketing 82 <https://journals.sagepub.com/doi/abs/10.1016/j.intmar.2016.12.002> accessed 16 May 2023.

99 Chris Jay Hoofnagle and Eduard Meleshinsky, 'Native Advertising and Endorsement: Schema, Source-Based Misleadingness, and Omission of Material Facts' (15 December 2015) <https://papers.ssrn.com/abstract=2703824> accessed 16 May 2023.

100 Chen Lou, Wenjuan Ma and Yang Feng, 'A Sponsorship Disclosure Is Not Enough? How Advertising Literacy Intervention Affects Consumer Reactions to Sponsored Influencer Posts' [2020] Journal of Promotion Management.

designed, taking into consideration user knowledge and the social and cultural contexts that shape user behaviour.

The user-facing initiatives also often suffer from the vice of information overload and user fatigue and put too much onus on individuals,[101] which has also rendered the e-Privacy Directive and the GDPR largely unsuccessful in creating meaningful transparency.[102]

Further, as discussed in the context of ad repositories, the ambiguous provision on disclosing "main parameters" for targeting can limit meaningful accountability, given advertisers often use methods beyond attributes to target users.[103] Further, even within the context of disclosing attribute information, platforms are most likely predisposed to disclose only limited targeting information in the presence of such ambiguous language. A study[104] on Facebook ad explanations shows how the platform only showed at most one attribute, irrespective of the number of attributes the advertisers chose. Further, their experiment revealed cases where the explanations mentioned an attribute as "may have been selected" when it was not selected by the advertiser, making the information not only incomplete but also misleading.

2.4. Transparency on How Platforms Profile Users

The singular focus on disclosure of "targeting parameters" provides some degree of transparency on the attributes employed by advertisers to target users. Still, it gives no transparency on how platforms ascribe those attributes to users or create categories for advertisers. Both in ad repositories and in user-facing disclaimers, the focus is on parameters for targeting and not how some of those parameters are derived from a variety of sources, including user behaviour and data brokers. While

101 Daniel Solove, 'The Limitations of Privacy Rights' (2023) 98 Notre Dame Law Review 975 <https://scholarship.law.nd.edu/ndlr/vol98/iss3/1>.

102 Leerssen and others (n 16).

103 Andreou and others (n 61).

104 ibid.

having transparency on advertisers' selection of targeting parameters is a significant step forward, it is equally important to understand how platforms classify users into interest groups, which can be used by political advertisers to target audiences.[105] This kind of classification often replicates existing power structures in society[106] and has downstream implications on how advertising, including in political campaigning, operates. For instance, it is important to examine how marginalised categories of gender, caste, ethnic or religious minority groups get encoded in targeting information offered by platforms to advertisers and how this impacts public discourse and democracy.

This transparency is all the more essential for the Global South, where platforms spend little resources,[107] have minimal understanding of local social contexts[108] and often do not have regional offices employing locals. The categorisation of audiences is determined by the platform's logic of economic value and profit and is inscrutable to users, researchers and regulators.[109] Consequently, there exists very little information on how microtargeting of ads plays out in the Global South. There has been little research on the harms that advertisement targeting can cause, both in terms of the distribution of economic opportunities as well as their impact on voter manipulation and offline violence.

---

105 Andreou refers to these two separate mechanisms as ad explanations (decisions of advertisers) and data explanations (decisions of platforms). See ibid.

106 See Rena Bivens and Oliver L Haimson, 'Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers' (2016) 2 Social Media + Society 2056305116672486 <https://doi.org/10.1177/2056305116672486> accessed 6 February 2024.

107 Takhshid (n 76); Ben Gilbert (n 52).

108 Paul Mozur, 'A Genocide Incited on Facebook, With Posts From Myanmar's Military' *The New York Times* (15 October 2018) <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html> accessed 1 May 2021; Jasper Jackson, Mark Townsend and Lucy Kassa, 'Facebook "Lets Vigilantes in Ethiopia Incite Ethnic Killing"' *The Observer* (20 February 2022) <https://www.theguardian.com/technology/2022/feb/20/facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing> accessed 5 June 2023; Giovanni De Gregorio and Nicole Stremlau, 'Inequalities and Content Moderation' (2023) 14 Global Policy 870 <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.13243> accessed 7 February 2024.

109 Kelley Cotter and others, '"Reach the Right People": The Politics of "Interests" in Facebook's Classification System for Ad Targeting' (2021) 8 Big Data & Society 2053951721996046 <https://doi.org/10.1177/2053951721996046> accessed 12 May 2023.

2.5. Meaningful User Control

The DSA provides an important step towards mandating some form of user control by providing an option to change the main targeting parameters.[110] However, various socio-political realities influence whether users understand targeting information and its implications for privacy and user experience. Adopting a purely technical approach to algorithmic accountability can have its limitations,[111] especially for users in the Global South who often have limited digital literacy and technical capacity.

Further, these users may share a complex relationship with platforms that might not always be characterised by distrust, even in the most adverse conditions. This is, for instance, reflected in research[112] on vulnerable users of exploitative instant loan platforms. It was found in the study that users often assumed responsibility for their negative experiences and viewed the platform services as aspirational. They attributed any negative experiences to their incompetence rather than unfair platform practices.[113] Thus, purely technical solutions won't be able to empower users with the agency to hold platforms accountable, as meaningful transparency requires a critical audience.[114] Moreover, individual users from marginalised communities might not be able to hold the platforms answerable even with the information made available to them given the platform-user power relations are an important determinant for platform accountability.[115]

---

110 DSA art 26(1)(d).

111 Nithya Sambasivan and others, 'Re-Imagining Algorithmic Fairness in India and Beyond', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) <https://dl.acm.org/doi/10.1145/3442188.3445896> accessed 17 May 2023.

112 Divya Ramesh and others, 'How Platform-User Power Relations Shape Algorithmic Accountability: A Case Study of Instant Loan Platforms and Financially Stressed Users in India', *2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) <https://dl.acm.org/doi/10.1145/3531146.3533237> accessed 16 May 2023.

113 ibid.

114 Jakko Kemper and Daan Kolkman, 'Transparent to Whom? No Algorithmic Accountability without a Critical Audience' (2019) 22 Information, Communication & Society 2081 <https://www.tandfonline.com/doi/full/10.1080/1369118X.2018.1477967> accessed 17 May 2023; Ananny and Crawford (n 79).

115 Ananny and Crawford (n 79).

Additionally, mediating factors like platform design and lack of information in local languages can prove to be significant impediments for many users in the Global South. Some studies [116] show that users rarely use the available voluntary advertisement controls due to a lack of awareness and instead prefer to use device application permission settings to control targeting by muting or unfollowing advertisement pages on social media.

The diversity of user experience also calls for more context-situated methods of user control that go beyond those provided in the DSA. Studies[117] have also shown how user perception of targeted advertisements varies across socio-cultural contexts. For instance, for those who use shared devices in common households, privacy from targeting can manifest as an option to not show targeted ads on such shared devices, especially the ads that the users might consider embarrassing or inappropriate for children.[118]

Thus, although advertisement disclaimers and options to change targeting parameters are good first steps in user-centric ad transparency, true user control would need a much more detailed disclosure that takes into account user experience and engagement. The information on what user action leads to a particular attribute being inferred for them could be truly enlightening for users and give them real control, as users often exercise choice through engaging with content rather than selecting abstract technical parameters.[119] This is all the more true for the Global South countries where there are many first-time internet users, and selecting abstract targeting parameters might be even less promising.

---

116 See Tanusree Sharma and others, 'User Perceptions and Experiences of Targeted Ads on Social Media Platforms: Learning from Bangladesh and India', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2023) <https://dl.acm.org/doi/10.1145/3544548.3581498> accessed 16 May 2023.

117 See Lalit Agarwal and others, 'Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising', *Proceedings of the Ninth Symposium on Usable Privacy and Security* (ACM 2013) <https://dl.acm.org/doi/10.1145/2501604.2501612> accessed 17 May 2023; Sharma and others (n 116).

118 ibid.

119 Leerssen (n 32); Andreou and others (n 61).

2.6. Complementary Legislation for Meaningful Transparency

Mandating ad archives and user-facing disclaimers can provide significant benefits to the Global South, but its objectives of meaningful transparency cannot be achieved in a vacuum. Identity verification for advertisers, especially political advertisers, is crucial. Many platforms voluntarily perform some baseline authorisation checks across several jurisdictions. However, these identity checks have limitations[120] as actors can use intermediaries or proxies to purchase ads for them and obfuscate the real source of financing. This can be used for corporate astroturfing or voter manipulation by malicious actors. Thus, complementing national legislation on electoral funding as well as regulatory oversight capable of enforcing the mandates on platforms is critical. These can prove to be limitations in those Global South countries where electoral legislation may have not yet grappled with the phenomena of online advertisement. Even when legislation exists, states may struggle with limited regulatory and law enforcement capacity to enforce compliance from platforms. Platforms also lack adequate human and automated resources trained in local contexts and languages. However, as a starting point in all countries, transparency legislation can require platforms to disclose how verification is done (if any).[121]

It is also important to note that data protection legislation is essential to prevent microtargeting harms and also to mitigate any privacy harms that might arise from the implementation of ad archives.[122]

120 Laura Edelson, Tobias Lauinger and Damon McCoy, 'A Security Analysis of the Facebook Ad Library', *2020 IEEE Symposium on Security and Privacy (SP)* (IEEE 2020) <https://ieeexplore.ieee.org/document/9152626/> accessed 3 May 2023.

121 Leerssen and others (n 16).

122 The archives must exclude any personal information for instance the custom targeting data that has contact information of users. Also, targeting data and user engagement data can lead to inferences on user demographics. See ibid.

Although meaningful transparency is important for platform accountability, it in itself is not a panacea to all the harms arising from advertising. Transparency is not an end in itself and should not be seen as an alternative to more binding regulation that might emerge from the research on ads, like banning behavioural targeting in advertisements.[123]

123 The DSA also prohibits targeting advertisements based on profiling children or based on special categories of personal information like sexual orientation or ethnicity. Paddy Leerssen (n 58).

Insights for the Global South

- Advertisement transparency is crucial for Global South countries. There is an urgent need to study how microtargeting of ads plays out in postcolonial societies with multiple social cleavages and younger political systems. There has been little research or understanding of the discrimination and harms that such practices cause, both in terms of the distribution of economic opportunities as well as their impact on voter manipulation and offline violence.

- Ad transparency can help raise general awareness and understanding of how ads operate and empower citizens to engage with questions of privacy, discrimination and fair and democratic elections.

- Mandating advertisement repositories comprising both commercial and political ads with detailed information on sponsors, financial spending, and targeting methods employed by advertisers, including targeting parameters, will be an important step forward from voluntary ad archives for the Global South. It is important to note that this additional transparency obligation is only applicable to VLOPs and VLOSEs under the DSA, as this might be a resource-intensive obligation for smaller platforms.

- User-facing disclaimers provide baseline transparency to users and can be useful to Global South users as well. However, more research should be undertaken to understand the efficacy of such disclaimers in different social, cultural, and economic contexts to design effective disclaimers for users with differing levels of digital literacy.

- Similarly, providing an option for users to control targeting parameters appears to be a good step forward. However, the real accountability derived from such a measure must be critically examined, and more holistic methods to provide meaningful control which goes beyond abstract technical parameters should be studied.

- Transparency on targeting parameters is a good step forward, however, any meaningful accountability from platforms would also need information on how platforms classify users into interest groups for advertisers.

- Limited state regulatory and enforcement capacity to monitor and audit the adequacy of information disclosed through these transparency mechanisms can be a limitation in the Global South. Further, platforms often raise jurisdictional issues, making it difficult for regulators and civil society actors to hold them accountable in local courts.

3. RISK MANAGEMENT

3.1. Introduction

The UN Guiding Principles for Business and Human Rights (UNGPs),1 endorsed by the UN Human Rights Council (UNHRC), have facilitated various forms of due diligence assessments of enterprises' impacts on fundamental human rights. Such assessments, where they focus on the risks that an enterprise or its systems can pose to human rights or values derived therefrom, are often termed 'human rights risk assessments'. Unlike audits, which are typically retrospective, risk assessments evaluate how an enterprise and its systems can prospectively impact human rights.2 Further, they are generally (and logically) followed by the implementation of appropriate safeguards to mitigate the identified risks.3

While obligations to undertake risk management (including risk assessment, mitigation and reporting) have become commonplace in environmental and health safety regulations,4 frameworks geared towards online harms and digital safety have

1 'Human Rights Reporting and Assurance Frameworks Initiative, 'UN Guiding Principles Reporting Framework', <https://shiftproject.org/resource/un-guiding-principles-reporting-framework/> accessed 30 May 2024.

2 Caitlyn Vogus and Emma Lanso, 'Making Transparency Meaningful: A Framework for Policymakers' (Centre for Comecracy and Technology, 2021) <https://cdt.org/wp-content/uploads/2021/12/12132021-CDT-Making-Transparency-Meaningful-A-Framework-for-Policymakers-final.pdf> accessed 30 May 2024. For differences between audits and risk assessments in the context of algorithmic systems, see Ada Lovelace Institute, 'Examining the Black Box: Tools for assessing algorithmic systems' (2020), <https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf>accessed 30 May 2024.

3 BSR, 'FAQ: Human Rights Assessment', <https://www.bsr.org/en/prs/human-rights-assessment> accessed 30 May 2024.

4 Zohar Efroni, 'The Digital Services Act: risk-based regulation of online platforms' (Internet Policy Review, 16 November 2021) <https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606> accessed January 23, 2025; Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 Harvard Law Review 524 <https://harvardlawreview.org/wp-content/uploads/2022/11/136-Harv.-L.-Rev.-526.pdf> accessed 30 May 2024.

not hitherto been required by law.[5] Most existing procedures for evaluating platforms' human rights impact have been implemented by major platforms either by themselves,[6] or advanced by civil society actors and researchers,[7] in response to growing public concern regarding the social, political, and economic risks posed by online services.[8]

The risk management mechanism set out under Articles 34 and 35 of the DSA thus represents a first-of-its-kind legislative intervention. It recognises that intermediaries categorised as Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), due to their wide reach and advertising-driven business models,[9] pose risks that operate at a societal scale.[10] Consequently, Article 34 requires them to assess risks stemming from the design, functioning, and use of

5 A noteworthy set of exceptions in this regard are requirements for data protection impact assessments, which require processors of personal data to assess the risks posed by their services on the protection of personal data. See, illustratively, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), art 35.

6 Human Rights Annual Report: Fiscal year 2021 (Microsoft 2021) <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE54vFs> accessed 30 May 2024; Meta, 'How Meta Is Preparing for Brazil's 2022 Election' (Meta, 12 August 2022) <https://about.fb.com/news/2022/08/how-meta-is-preparing-for-brazils-2022-elections/> accessed 30 May 2024; Meta, 'Meeting the Unique Challenges of the 2020 Elections' (Meta, June 26 August 2020) <https://about.fb.com/news/2020/06/meeting-unique-elections-challenges/>accessed 30 May 2024.

7 BSR, 'Human Rights Assessment: Global Internet Forum to Counter Terrorism' (2021) <https://gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf> accessed 30 May 2024; Danish Institute of Human Rights, 'Guidance on Human Rights Impact Assessment of Digital Activities' (2020) <https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities> accessed 30 May 2024; Global Network Initiative, 'GNI Assessment Toolkit' (October 2021) <https://globalnetworkinitiative.org/wp-content/uploads/2021/11/AT2021.pdf> accessed 30 May 2024.

8 See Efroni (n 4), which discusses the increasing role of the concept of 'risk' in regulations surrounding the digital economy and society, including the EU's regulations on artificial intelligence and data protection.

9 DSA 2022, recital 79.

10 See Efroni (n 4).

their services in the European Union (EU). The assessment must include consideration of certain identified categories of 'systemic risks':[11]

(a) the dissemination of illegal content; and
(b) negative effects on:
    1) the exercise of fundamental rights under the Charter of Fundamental Rights of the EU (EU Charter);[12]
    2) civic discourse and electoral processes;
    3) public security; and
    4) gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being.

In assessing such risks, VLOPs and VLOSEs must specifically consider the influence of the following factors:[13]

(a) the design of their recommender systems and other algorithmic systems;
(b) their content moderation systems;
(c) their Terms and Conditions (T&Cs) and their enforcement;
(d) their systems for selecting and presenting advertisements; and
(e) their data-related practices.

Further, the assessment must analyse whether and how each of the risks identified above may be affected by the intentional manipulation of their services, including by inauthentic use (such as through fake accounts or stolen identities) or automated exploitation (such as through coordinated social bots), and amplification and wide dissemination of content that is illegal or violative of their T&Cs.

11 DSA 2022, art 34(1).

12 *Charter of Fundamental Rights of the European Union* [2012] OJ C326/391 <https://www.europarl.europa.eu/charter/pdf/text_en.pdf> accessed 22 January 2025.

13 DSA 2022, art 34(2).

Such assessments must be conducted at least once every year. Additionally, they must be conducted before the VLOP or VLOSE deploys any functionality likely to have a 'critical impact' on such risks.[14]

Once systemic risks have been assessed with reference to their severity and their probability, VLOPs and VLOSEs must institute effective measures to mitigate them.[15] These measures can include, illustratively, adapting their online interfaces, T&Cs or algorithmic systems, enhancing user-transparency, enhancing cooperation with other intermediaries or trusted flaggers, or changing their internal procedures.[16] These measures must be reasonable, proportionate, tailored to the risks identified, and must reflect particular regard for fundamental rights.[17]

The duty to effectively conduct such risk management rests, amongst other due diligence obligations under the DSA, with the 'compliance function' of the VLOP/VLOSE. This division is required to be independent from other operational functions of the VLOP/VLOSE.[18] Further, to ensure that the results of such risk management effectively facilitate investigation and inform future assessments,[19] the assessment and mitigation reports must be preserved for at least three years.[20] Further, redacted versions of such reports must be submitted to regulatory authorities and made publicly available every year.[21] Such submissions will form the basis of an annual report published by the EC, identifying the most prominent and recurring risks arising from VLOPs' and VLOSEs' services, and the best practices towards their mitigation.[22]

14 DSA 2022, art 34(1).

15 DSA 2022, art 35(1).

16 DSA 2022, art 35(1)(a)-(k).

17 DSA 2022, art 35(1).

18 DSA 2022, art 41(1).

19 DSA 2022, recital 85.

20 DSA 2022, art 34(3).

21 DSA 2022, art 42(4).

22 DSA 2022, art 35(2).

3.2. Risk Identification and Assessment

The risk assessment mechanism under the DSA ostensibly exhibits a shift towards a systems-approach to platform regulation,[23] by requiring VLOPs and VLOSEs to *ex ante* assess risks considered 'systemic'. This shift from individual outcomes to broader procedures,[24] has been viewed as a progression from existing regulatory approaches, limited by their disproportionate focus on *ex post* affixation of liability on platforms for individual pieces of content.[25] It is expected to at least prompt platforms to consider their operational risks proactively and methodically, instead of reacting to harms as they magnify.[26] If platforms were to similarly assess risks posed by their services in Global South jurisdictions, the insights gained could significantly assist regulators, public stakeholders, and platforms themselves, in formulating responses to address such risks.

However, the primary task of legislatively stipulating the particular risks that platforms must monitor is far from simple.[27] Foremost, any meaningful framework for the assessment of risks arising out of platforms' services derives its legitimacy from a robust normative foundation. The EU Charter, which guarantees certain inalienable rights to individuals across the EU, provides such a foundation for risk assessments under the DSA. In fact, the DSA and its risk management framework are expressly geared towards protecting the rights enshrined in the EU Charter, alongside certain other societal interests and values.[28] As a corollary, the EU Charter and its interpretation are also expected to govern the implementation of the framework. [29] While most Global South states have signed and ratified the

23 Douek (n 4).

24 Daphne Keller, 'The DSA's Industrial Model for Content Moderation' (*Verfassungsblog*, 24 February 2022) <https://verfassungsblog.de/dsa-industrial-model/> accessed 30 May 2024.

25 Douek (n 4).

26 Douek (n 4).

27 Douek (n 4).

28 See, for instance DSA art 1(1), 34(1)(b) and 35(c); DSA recitals 153 and 155.

29 Eliška Pírková and others, 'Towards Meaningful Fundamental Rights Impact Assessments under the DSA', Access Now and European Centre for Not-for-Profit Law (September 2023)

International Covenant on Civil and Political Rights,[30] the extent to which they incorporate human rights protections into their constitutional documents or national laws varies significantly.[31] For such states, the absence of expressly enumerated human rights could pose an obstacle to the identification of suitable and legitimate risk-categories.

Even in jurisdictions where human rights are expressly guaranteed, the contextual and dynamic nature of the risks posed by platforms' services represents a significant challenge to their identification.[32] Such risks relate not only to the nature of a platform's services, but also to how these services interact with a particular social, economic and political environment, at a particular time.[33] Their severity and likelihood hinge substantially on the broader information ecosystem, shaped by a variety of factors – for instance, the popularity of the platform's services amongst various user-groups, the presence of viable alternatives and the state's influence on both traditional and online media. Moreover, since platforms have historically withheld access to the data required for public interest research, external stakeholders have struggled to observe causal linkages between their services and

---

<https://www.accessnow.org/wp-content/uploads/2023/09/DSA-FRIA-joint-policy-paper-September-2023.pdf.> accessed 30 May 2024.

30 *International Covenant on Civil and Political Rights* (adopted 16 December 1966 UNGA Res 2200A (XXI)) <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights> accessed 22 January 2025; ; UN Human Rights: Office of the High Commissioner, 'Status of Ratification Status Interactive Dashboard – International Covenant on Civil and Political Rights' <https://indicators.ohchr.org/> accessed 22 January 2025.

31 Zachary Elkins and others, 'Getting to Rights: Treaty Ratification, Constitutional Convergence, and Human Rights Practice' (2013) 54 Harvard International Law Journal 61.

32 Zohar Efroni, 'The Digital Services Act: risk-based regulation of online platforms' [2021] Internet Policy Review <https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606> accessed 30 May 2024; Robin Mansell and others, 'Information Ecosystems and Troubled Democracy: A Global Synthesis of the State of Knowledge on News Media, AI and Data Governance' Observatory on Information and Democracy (January 2025) <https://observatory.informationdemocracy.org/wp-content/uploads/2024/12/rapport_forum_information_democracy_2025.pdf> accessed 22 January 2025.

33 See Efroni (n 4).

their adverse societal effects.[34] Such linkages have been established only retrospectively, as they accumulate over time.[35] In these circumstances, it may be enormously difficult to foresee and enumerate the various kinds of risks that can arise from platforms' services in a particular jurisdiction.

The challenge has been evident in the context of the DSA as well. While Article 34 identifies an expansive set of risks, the inclusion of certain societal risks alongside risks to particular rights under the EU Charter), has drawn criticism. Some apprehend that the wide variety and the breadth of such risks would hinder the development of targeted assessment tools and methods for specific risks, such as algorithmic bias or coordinated influence operations.[36] Others note that the prioritisation of certain human rights over others ignores their mutually affirming character, as underlined by the UNGPs.[37] Further, certain categories of risks, such as "the dissemination of illegal content" and "negative effects on electoral practices", leave enormous room for interpretation,[38] and ambiguity regarding how they relate to the human rights framework.[39] As Barata observes, such ambiguous risk-

34 AlgorithmWatch and others, 'DSA must empower public interest research with public data access' AlgorithmWatch (31 May 2023) <https://algorithmwatch.org/en/dsa-empower-public-interest-research-data-access/> accessed 22 January 2025.

35 Perkova and others (n 29). Efroni (n 4)

36 Alessandro Mantelero, 'Fundamental rights impact assessments in the DSA' (*Verfassungsblog*, 1 November 2022, <https://verfassungsblog.de/dsa-impact-assessment/> accessed 30 May 2024; Paddy Leerssen, 'Counting the days: what to expect from risk assessments and audits under the DSA – and when?' (*DSA Observatory*, 30 January 2023) <https://dsa-observatory.eu/2023/01/30/counting-the-days-what-to-expect-from-risk-assessments-and-audits-under-the-dsa-and-when/> accessed 30 May 2024.

37 'How can we apply human rights due diligence standards to content moderation? Focus on the EU Digital Services Act', Centre for Democracy and Technology (29 July 2021) <https://globalnetworkinitiative.org/wp-content/uploads/2021/09/CDT-GNI-DSA-Due-Dilligence-July-29.pdf>

38 The DSA reserves powers for the EC to formulate delegated legislation to detail or clarify the scope of many of its provisions. Notably however, it does not expressly reserve any such power for the EC to clarify the scope of the statutory risk-parameters set out under Section 34.

39 Iverna McGowan and Ashal Allen, 'Fostering responsible business conduct in the tech sector – the need for aligning risk assessment, transparency and stakeholder engagement provisions under the EU Digital Services Act with the UNGPs', Centre for Democracy & Technology (24 August 2023) <https://cdt.org/insights/fostering-responsible-business-conduct-in-the-tech-sector-the-need-for-

formulations may nudge platforms to restrict or demote "otherwise legal borderline content", in attempts to mitigate the associated risks.[40] Moreover, since such visibility-reductions would be carried out at a systemic level, users and other public stakeholders would be hard-pressed to identify and seek accountability for them. Opaque restrictions of this nature, based on perceived risks that do not have clear linkages to protected human rights, could endanger users' freedom of speech and expression, as well as their right to receive information.

Thus, any jurisdiction attempting to institute a risk assessment mechanism must navigate the challenge of identifying the categories of risks that platforms could pose in that jurisdiction. Considering the socio-economic heterogeneity of users even within a Global South jurisdiction, drawing up a definitive list of 'at-risk' rights and values, tailored to the jurisdiction may be particularly difficult – especially without adequate empirical evidence to demonstrate the effects of platforms' services on such rights and values, as noted earlier. Further, ambiguous formulations, such as risks to 'public security' and 'civic discourse' could be interpreted expansively by authoritarian governments, as a pretext for curtailing users' rights to free speech and information.[41]

Even where systemic risks are suitably identified, effective assessment of their severity and probability in a particular Global South jurisdiction would demand significant cultural and linguistic expertise.[42] In the *status quo*, most major platforms are disproportionately staffed by personnel from the Global North and are

---

aligning-risk-assessment-transparency-and-stakeholder-engagement-provisions-under-the-eu-digital-services-act-with-the/> accessed 30 May 2024.

40 Joan Barata, 'The Digital Services Act and its impact on the right to freedom of expression: Special focus on risk mitigation obligations', [2021] Plataforma por la Libertafd de Informacion<https://libertadinformacion.cc/wp-content/uploads/2021/06/DSA-AND-ITS-IMPACT-ON-FREEDOM-OF-EXPRESSION-JOAN-BARATA-PDLI.pdf> accessed 30 May 2024; Efroni (n 4).

41 See Anupam Chander, 'When the Digital Services Act Goes Global' <https://scholarship.law.georgetown.edu/facpub/2548/> accessed 30 May 2024; Centre for Communication Governance, LIRNEAsia and BRAC University, 'Social Media Regulation and the Rule of Law: Key Trends in Sri Lanka, India and Bangladesh' CLJ Malaysia Sdn. Bhd. (2024) <https://www.kas.de/en/web/rspa/single-title/-/content/publication-social-media-platforms-regulation-and-the-rule-of-law> accessed 22 January 2025.

42 Douek (n 4); Mantelero (n 36).

unlikely to have such expertise readily available.[43] To facilitate context-sensitive assessments of risks in Global South jurisdictions, it is critical that such platforms diversify the compositions of their staffs as well as engage extensively with civil society and academic actors familiar with the relevant local contexts.

The above discussion suggests that identifying and assessing categories of risks in a principles-based and contextual manner, while also preserving legal certainty for compliance, is a fraught venture, normatively as well as practically. Even so, to enable deeper analysis of how various risks relate to one another as well as how they impact user-groups differentially, Asha Allen advocates that risk assessments follow an 'intersectional' approach.[44] Methodologically, this would require platforms to inquire how a set of risks intersects with existing hierarchies. For example, any assessment of risks associated with online gender-based violence would include an assessment of how such risk is magnified for persons belonging to historically marginalised groups. Further, this would also entail that platforms consider how one category of risks interacts with another. As an illustration, any assessment of the health risks arising from the dissemination of images depicting violence during an armed conflict, would also consider how blocking access to such images would affect users' right to receive information regarding the conflict. Such an approach would facilitate a more meaningful comprehension of the way risks arising out of online interactions play out in specific contexts, instead of analysing them in silos. It may be particularly advisable in the Global South, where long-enduring power structures

43 Kalev Leetaru, 'The Importance of Context and Intent in Content Moderation' Forbes (28 July 2019) <https://www.forbes.com/sites/kalevleetaru/2019/07/28/the-importance-of-context-and-intent-in-content-moderation/?sh=73aed1852a95> accessed 30 May 2024; Farhana Shahid and Aditya Vashistha, 'Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?' Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems <https://www.adityavashistha.com/uploads/2/0/8/0/20800650/decolonial-chi-2023.pdf> accessed 30 May 2024.

44 Asha Allen, 'Ann Intersectional Lens on Online Gender Based Violence and the Digital Services Act' (*Verfassungsblog*, 1 November 2022) <https://verfassungsblog.de/dsa-intersectional/> accessed 30 May 2024; See Kimberle Crenshaw, 'Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color' 43(6) Stanford Law Review 1241-1299 <https://blogs.law.columbia.edu/critique1313/files/2020/02/1229039.pdf> accessed 30 May 2024 .

and social hierarchies (such as those based on class, caste, race, and gender) often impact online discourse acutely.

3.3. Risk Mitigation

Following the assessment of systemic risks, the DSA requires VLOPs and VLOSEs to institute measures to mitigate such risks. It illustrates a broad range of mitigation measures that they can implement.[45] Some of these relate to the design of their services, such as their online interfaces,[46] or their algorithmic systems.[47] Others relate to the processes adopted by them, such as those for content moderation[48] or for risk identification.[49] Further, certain other listed measures relate to cooperation with external stakeholders, such as trusted flaggers [50] or other online intermediaries.[51]

While this illustrative list will guide VLOPs and VLOSEs in formulating responses to address identified risks, key questions remain. Crucially, the DSA does not stipulate either the threshold of risk at which such measures should be implemented or the threshold that such measures must meet to be considered adequate. It only directs VLOPs/VLOSEs to employ common risk parameters, such as severity, probability, likelihood, scale and reversibility.[52]

45 DSA 2022, art 35(1).

46 DSA 2022, art 35(1)(a).

47 DSA 2022, art 35(1)(d).

48 DSA 2022, art 35(1)(c).

49 DSA 2022, art 35(1)(f).

50 DSA 2022, art 35(1)(g).

51 DSA 2022, art 35(1)(h).

52 DSA 2022, recital 56. Pertinently, the delegated regulation on audits under the DSA, finalised in October 2023, sheds more light on certain specific aspects that auditors must examine, when checking for an intermediary's compliance with risk management obligations. These include, for instance, examining the sources of information used for identification of risks; and whether the measures undertaken for risk mitigation respond collectively to all the risks identified.

The lack of prescriptive clarity on the suitability and adequacy of risk mitigation measures is a consequence of the inherent difficulty in tying risk management to substantive outcomes. As Douek notes, empirically assessing or demonstrating the 'impact' of specific measures to regulate online content is highly contentious.[53] Risk management frameworks can, at best, compel platforms to exhibit foresight of the risks that their services can pose and to demonstrate that they have certain procedural safeguards in place to address them. They can require platforms to justify whether and how they synthesise the risks arising from their services (say, dissemination of inflammatory but legal content) with other competing considerations (users' rights to free speech). Unlike outcome-based mechanisms (such as notice-and-action obligations, which require platforms to remove illegal content upon receiving actual knowledge), these procedural frameworks cannot be used to impose specific measures on platforms, and affix liability upon their failure to implement them. So long as platforms can demonstrate that certain mitigation measures have been instituted, they retain broad discretion to decide the optimal responses to the identified risks at any instance. Requirements such as those of reasonability, proportionality and effectiveness laid down under the DSA,[54] can only provide indicative guidance for platforms in formulating such responses, and for auditors in retrospectively evaluating them.

The description of risks in the language of "any negative effects (on human rights)" under the DSA, may add to the indeterminacy of mitigation obligations. It seems to downplay the competing nature of rights and the frequent need to balance or harmonise them in the given context – a task typically undertaken by judicial authorities, on the basis of arguments put forth by litigants and relying on years of jurisprudence.[55]

53 Douek (n 4).

54 DSA 2022, art 37(1).

55 Barata (n 40).

Such broad discretion and lack of 'enforceability' of measures, strikes at the value of risk management as an accountability mechanism. As a result, risk management frameworks may do little to displace platforms from their positions as *de facto* gatekeepers of online speech. On the contrary, entrusting them to prioritise between competing and often-conflicting human rights and select the appropriate mitigation measures, could potentially cement their gatekeeping positions.[56]

Any Global South jurisdiction attempting to impose risk management obligations must contend with the above limitations and challenges. Additionally, as noted in the context of independent audits, the limitations in Global South states' regulatory capacities (both technical and financial) and their bargaining powers vis-à-vis major platforms may further affect their ability to impose and effectively oversee mitigation measures.[57]

Considering the complexity and the politically contentious nature of risk mitigation measures, ensuring the meaningful engagement of a diverse range of stakeholders (including those affected disproportionately by the risks) in their formulation is crucial.[58] The DSA creditably does suggest avenues for such engagement. It advises VLOPs and VLOSEs to consult representatives of users, of groups particularly affected by their services, independent experts and civil society, in formulating mitigation measures.[59] Further, it recommends that inputs from such consultations be fed into risk management methodologies.[60] Such engagement could be facilitated

56 ibid.

57 Zahra Takhshid, 'Regulating Social Media in the Global South', 24 Vanderbilt Journal of Entertainment and Technology Law 1 (2022), <https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1564&context=jetlaw> accessed 30 May 2024.

58 Mozilla position paper on the EU Digital Services Act, *Mozilla* (May 2021) <https://blog.mozilla.org/netpolicy/files/2021/05/Mozilla-DSA-position-paper-.pdf> accessed 30 May 2024; McGowan and Allen (n 39); Perkova and others (n 29).

59 DSA 2022, recital 89.

60 DSA 2022, recital 90.

through discussions on voluntary codes of conduct for risk management.[61] As researchers at the Centre for Democracy & Technology observe, assessments of this nature would require expertise in both social science and computer science, to engage with the range of social and technical issues at stake.[62]AccessNow and the European Center for Not-for-profit Law (ECNL) have gone a step further, to recommend the engagement of experts in engineering, product development, research, risk management, legal, policy, finance, sustainability, communications, marketing, sales, human resources, trust and safety, and human rights.[63] If a Global South jurisdiction decides to impose a risk management mechanism by law, it may be advisable for requirements of multi-stakeholder engagement to be embedded into risk management frameworks, so that risk management takes the character of a truly public exercise. The metrics for stakeholder engagement, proposed by AccessNow and ECNL, offer valuable guidance for platforms to conduct such engagement, and for other stakeholders to evaluate them.[64]

Further, to cover existing gaps in benchmarks and methodologies for risk management,[65] it is critical to accelerate empirical research regarding the categories and evolution of risks posed by platforms' services, particularly in the Global South. In undertaking such research, other transparency mechanisms such as independent audits and providing researchers access to data held by platforms, can be particularly beneficial.[66] Such mechanisms would offer insights to guide platforms as well as regulators in framing effective and context-sensitive responses to risks.

61 DSA 2022, art 45(2).

62 McGowan and Allen (n 39).

63 Perkova and others (n 29).

64 ibid.

65 Mantelero (n 36).

66 Mozilla (n 58); Barata (n 40).

3.4. Reporting

The DSA requires reports of risk assessment conducted as well as risk mitigation measures adopted, to be made publicly available.[67] As noted in the context of algorithmic risk management,[68] the publication of platforms' decisions and the rationale behind them would enable future review, create leverage for regulatory measures and enhance procedural accountability. Further, much like audit reports, such reports would provide other platforms reference points and encourage the development of industry best-practices and standards.[69] Without doubt, there are currently numerous gaps in public and regulatory understanding of risks posed by platforms' services and ways to mitigate them; in this context, information gained from successive risk management cycles can assist in plugging these gaps and lead to the iterative refinement of risk management mechanisms. Further, the aggregation of reports from various VLOPs and VLOSEs will enable holistic analyses of risks at the ecosystem level, in addition to risks generated by individual services. Additionally, such information will make other interlinked oversight mechanisms, such as independent audits, more robust.[70]

Crucially, much of this will depend on the extent to which VLOPs and VLOSEs actually disclose information in their public reports. Like other transparency mechanisms under the DSA, the mechanism for risk management reporting also allows a VLOP/VLOSE to redact information from public reports on certain grounds – if disclosure of such information would result in disclosure of confidential information, cause significant security vulnerabilities for its service, undermine

67 DSA 2022, art 42(4).

68 Andrew D. Selbst, 'An Institutional View of Algorithmic Impact Assessments', 35 Harvard Journal of Law & Technology 117 (2021) <https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf> accessed 30 May 2024.

69 Douek (n 4).

70 Mozilla (n 58); Claire Pershan, 'Cutting Through the Jargon - Independent Audits in the Digital Services Act' *Mozilla* (30 January 2023) <https://foundation.mozilla.org/en/blog/cutting-through-the-jargon-independent-audits-in-the-digital-services-act/> accessed 30 May 2024; Nicolo Zingales, 'The DSA as a paradigm shift for online intermediaries' due diligence' (*Verfassungsblog*, 2 November 2022)<https://verfassungsblog.de/dsa-meta-regulation/> accessed 30 May 2024.

public security or harm users.[71] While any redaction has to be justified with a statement of reasons, the breadth of the grounds allows intermediaries significant flexibility to limit their public disclosures.[72] To maximise the transparency gains discussed above, all such redactions to public versions of risk assessment and mitigation reports must be grounded in principles of reasonability and proportionality.

If a Global South jurisdiction institutes similar reporting obligations on online platforms regarding their risk management procedures, it must ensure that any grounds for redaction of information are narrowly and reasonably framed. Further, to derive meaningful lessons and maximise the transparency gains from such reports, it must address constraints in regulatory capacity and remove barriers to independent research, regarding risks arising from platforms' services in that jurisdiction. Without "polycentric oversight"[73] enabled by such measures, regulatory risk management procedures may end up resembling perfunctory safety assessments conducted behind closed doors by platforms.

71 DSA, art 42(5).

72 Nayanatara Ranganathan, 'Regulating influence, timidly' (*Verfassungsblog*, 4 November 2022) <https://verfassungsblog.de/dsa-regulating-influence/> accessed 30 May 2024.

73 Mozilla (n 58)

Insights for the Global South

- The risk management framework under the DSA exhibits a shift towards an *ex ante* systemic approach to intermediary regulation, where VLOPs and VLOSEs are required to pre-emptively and periodically assess the potential societal risks that may arise from the use of their services.

- The framework is expected to prompt platforms to consider their operational risks proactively and methodically, instead of reacting to harms as they magnify. If platforms were to similarly evaluate risks posed by their services in Global South jurisdictions, the insights gained could significantly assist regulators, public stakeholders, and platforms themselves, in formulating responses to address such risks.

- Specifying and encoding the categories of societal risks that platforms should assess can be a thorny task. Such risk-categories must be firmly tethered to values that have legal as well as normative acceptance, in the particular jurisdiction. Global South states that have not instituted a human rights framework in their constitutional documents or domestic law must contemplate alternative frameworks, to ground any risk management obligations that they seek to impose on platforms.

- The dynamic and contextual nature of societal risks posed by online platforms represents another challenge to the identification of risk-categories in law. On one hand, broad formulations of risk-categories may be difficult for platforms to assess, and would nudge them to restrict or demote borderline-legal content to comply with their obligations, particularly in the absence of the relevant cultural and linguistic expertise. On the other hand, highly prescriptive formulations can result in a framework that becomes anachronistic with changes in the socio-political context.

- Any Global South jurisdiction considering a risk assessment framework must promote rigorous research to identify the categories of risks that online platforms pose in that jurisdiction, how such risks evolve over time, and how

such risks intersect with social, political and economic structures and processes in that jurisdiction.

- In addition to the identification of risk-categories, it is important to develop suitable benchmarks for mitigation of risks. Such benchmarks should at least offer guidance regarding the threshold of risk at which mitigation measures should be implemented, and the threshold that such measures must meet to be considered adequate. However, it is important to understand that unlike outcome-based mechanisms (like notice-and-action), procedural frameworks like risk assessments are limited in their "enforceability" and cannot be tied to particular outcomes.

- Considering the complexity and the politically contentious nature of risk mitigation measures, a diverse range of stakeholders (including those affected disproportionately by the risks) must be meaningfully engaged in their formulation. Any Global South jurisdiction contemplating a risk management mechanism should consider making such engagement mandatory.

- Intermediaries must report the results of the risk assessments conducted and mitigation measures adopted, to the general public. Such reports would provide other intermediaries reference-points and encourage the development of industry best-practices and standards. To maximise the transparency gains from such reports, all redactions to public versions of risk assessment and mitigation reports must be grounded in principles of reasonability and proportionality.

4. AUDITS

4.1. Introduction

Typically, the term "audit" is used to describe a retrospective assessment, focussed on evaluating an entity's compliance with a predetermined set of standards within a particular period.[1] Audits are designed to serve two interlinked purposes: providing assurance regarding an entity's business practices against certain standards, and identifying gaps or areas where the entity fails to meet them.

Audits can be internal, where the entity assesses itself, or external, where an external functionary conducts the assessment, usually with the entity's cooperation. Further, they may either be mandatory (under an applicable law, a code of conduct or an agreement); or conducted voluntarily by an entity, to signal transparency and demonstrate compliance with commonly accepted standards that are relevant to its operations.[2]

While audits and assessments aimed at verifying compliance with financial regulations have a long history,[3] those aimed at assessing businesses' impacts on

1 Caitlyn Vogus and Emma Lanso, 'Making Transparency Meaningful: A Framework for Policymakers' (Centre for Democracy and Technology, 2021) <https://cdt.org/wp-content/uploads/2021/12/12132021-CDT-Making-Transparency-Meaningful-A-Framework-for-Policymakers-final.pdf> accessed 30 May 2024. For differences between audits and risk assessments in the context of algorithmic systems, see Ada Lovelace Institute, 'Examining the Black Box: Tools for assessing algorithmic systems' (2020), <https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf> accessed 30 May 2024.

2 See Digital Regulation Cooperation Forum, 'Auditing algorithms: the existing landscape, role of regulators and future outlook' (September 2022), <https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook#:~:text=Algorithmic%20auditing%20refers%20to%20a,to%20inspecting%20its%20inner%20workings> accessed 30 May 2024.

3 Vogus and Lanso (n 1).

human rights have emerged only in recent decades.[4] The UN Guiding Principles on Business and Human Rights, endorsed by the UNHRC,[5] and the OECD Guidelines for Multinational Enterprises,[6] represent important sets of criteria for such assessments.[7] Recognising platforms' profound impact on the human rights to free speech, information and privacy, numerous multi-stakeholder coalitions have built upon the Guiding Principles to formulate benchmarks and conduct systematic assessments – notable examples include the Electronic Frontier Foundation's Who Has Your Back assessment (initiated in 2011),[8] the GNI Company Assessments (initiated in 2013)[9] and the Ranking Digital Rights Corporate Accountability Index (initiated in 2015).[10] The B-Tech Project, launched in 2019 by the UN, provides indicative guidance for operationalising the Guiding Principles in the technology sector through concrete policy recommendations.[11] Additionally, certain platforms have also conducted or submitted themselves to audits, to reassure stakeholders regarding their content moderation practices and their impact on human rights – Meta's Civil Rights Audit (from 2018 to 2020), conducted by civil rights activist

4 Notably, since the introduction of the GDPR, audits for the limited purpose of ensuring compliance with data protection regulation, have gained prominence. However, data protection audits are typically conducted internally by companies and their results are not disclosed to the general public.

5 'Human Rights Reporting and Assurance Frameworks Initiative, 'UN Guiding Principles Reporting Framework', <https://shiftproject.org/resource/un-guiding-principles-reporting-framework/> accessed 30 May 2024.

6 OECD Guidelines for Multinational Enterprises on Responsible Business Conduct (2023), <https://mneguidelines.oecd.org/mneguidelines/> accessed 30 May 2024.

7 United Nations Human Rights Office of the High Commissioner, 'Guiding Principles on Business and Human Rights' (2011) <https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf> accessed 30 May 2024.

8 Gennie Gebhart, 'Who Has Your Back? Censorship Edition 2019' (*EFF*, June 12, 2019), <https://www.eff.org/wp/who-has-your-back-2019> accessed 30 May 2024.

9 Global Network Initiative, 'GNI Assessment Toolkit' (October 2021) <https://globalnetworkinitiative.org/wp-content/uploads/2021/11/AT2021.pdf> accessed 30 May 2024.

10 Ranking Digital Rights, 'The 2020 RDR Index' (2020) <https://rankingdigitalrights.org/index2020/> accessed 30 May 2024.

11 UNOHCHR 'B-Tech Project: OHCHR and business and human rights', <https://www.ohchr.org/en/business-and-human-rights/b-tech-project> accessed 30 May 2024.

Laura Murphy at the behest of civil society and members of the U.S. Congress, provides perhaps the foremost example.[12]

Against this background, the adoption of audits under Article 37 of the DSA is a significant step, since it mandatorily requires each Very Large Online Platform (VLOP) and Very Large Online Search Engine (VLOSE) to commission an audit at least once every year, at its own cost. Such a "second-party" audit, i.e. by external auditors appointed by the VLOP/VLOSE, must verify the VLOP/VLOSE's compliance with each of its due diligence obligations under the DSA, including the risk assessment and mitigation obligations detailed in Chapter 3 (*Risk Management*). Further, the audit must verify their fulfilment of any commitments undertaken pursuant to any voluntary code of conduct and/or any crisis protocol. To ensure that audits are conducted efficiently and effectively, the VLOP/VLOSE must provide the necessary cooperation and assistance to auditors, by giving them access to all relevant data and premises and answering oral or written questions.[13] Additionally, auditors can access other reliable sources of relevant information, including studies conducted by vetted researchers, underlining the interlinked nature of the DSA's transparency mechanisms.[14] Auditors, on their part, must ensure an adequate level of confidentiality and professional secrecy, regarding the information obtained in the context of the audits, even after such audits have been concluded.[15]

To safeguard the credibility of audits, Article 37 prescribes certain eligibility criteria for the appointment of organisations as auditors. Such criteria relate, first, to their

12 Laura Murphy and others, 'Facebook's Civil Rights Audit - Final Report' (*Facebook*, 8 July 2020) <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf> accessed 30 May 2024.

13 DSA art 37(2); DSA recital 92.

14 See DSA art 40(4); Claire Pershan, 'Cutting Through the Jargon - Independent Audits in the Digital Services Act' Mozilla (30 January 2023) <https://foundation.mozilla.org/en/blog/cutting-through-the-jargon-independent-audits-in-the-digital-services-act/> accessed 30 May 2024.

15 DSA 2022, art 37(2).

financial and professional independence from the VLOP/VLOSE that they are to audit; and second, to their technical and professional competence.[16]

At the conclusion of each audit, the auditor(s) must prepare an audit report, containing, *inter alia*, a description of the elements audited, the methodology applied, the summary of main findings and the third parties consulted for the audit.[17] Crucially, the report must specifically contain an audit opinion ('positive', 'positive with comments' or 'negative') on whether the VLOP/VLOSE complied with each applicable obligation and commitment.[18] Where the opinion is not 'positive' for any obligation, the report should provide operational recommendations to achieve compliance, with recommended time-frames.[19] Further, where the auditors fail to audit certain elements that should fall within the scope of the audit or fail to reach a conclusion regarding any such elements, the report should specifically explain the reasons for such failure.[20]

Within a month of receiving such recommendations, the VLOP/VLOSE must adopt an audit implementation report, setting out the measures taken to implement the recommendations. [21] If it does not implement any such recommendation, the VLOP/VLOSE must justify reasons for not doing so and set out any alternative measures taken to address the identified non-compliance.[22]

Upon completion of the audit, the audit report and the audit implementation report (where relevant) must be submitted to regulatory bodies. Further, within three months of the receipt of the audit report, both the audit report and the implementation report must be made publicly available by the VLOP/VLOSE. Notably, the VLOP/VLOSE may redact certain information from the public versions

16 DSA 2022, art 37(3).

17 DSA 2022, art 37(4)(a) to (d).

18 DSA 2022, art 37(4)(g).

19 DSA 2022, art 37(4)(h).

20 DSA 2022, art 37(5), recital 93.

21 DSA 2022, art 37(6).

22 DSA 2022, art 37(6).

of these reports, in the interest of confidentiality (for itself or for its users), for the security of its service, to protect public security or to avoid harm to users.[23] Any such removal must be justified with a statement of reasons.[24]

4.2. Audit as a Tool for Transparency

While voluntary audits and assessments conducted have contributed to public understanding of platforms' functioning, their efficacy has been limited – this is partly due to their reliance on voluntary cooperation by platforms and the limited information made available for such assessments.[25] In contrast, Article 37 obligates VLOPs and VLOSEs, by law, to disclose information on their designs, policies, and procedures to external experts. Further, auditors are empowered to conduct additional inquiries and investigations to gather the information necessary for verifying compliance, with compulsory cooperation by VLOPs and VLOSEs. A substantial portion of such information can be expected to find its way into the final audit reports prepared by them.

For regulators, such information can be valuable, not only as an accountability tool, but as an evidence-base to iteratively inform platform regulation. For users, researchers, and other stakeholders, such reports, where published with suitable levels of detail and complexity, can provide a "comparative basis for public scrutiny"[26] of VLOPs and VLOSEs. They can spur further discourse on platforms' impact on the meaningful exercise of the rights to free speech, access to information,

23 DSA 2022, arts 42(5), 37(2).

24 DSA 2022, art 42(5).

25 Ben Wagner and Lubos Kuklis, 'Establishing Auditing Intermediaries to Verify Platform Data' in Martin Moore and Damian Tambini (eds), *Regulating Big Tech: Policy Responses to Digital Dominance* (New York, 2021, Oxford Academic, October 2021) <https://academic.oup.com/book/39213/chapter/338717733> accessed 30 May 2024; Vogus and Lanso (n 1).

26 European Commission, 'Delegated Regulation on independent audits under the Digital Services Act' (2023), <https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act> accessed 30 May 2024.

and privacy.[27] Information gained from an audit report can also guide other platforms (including those not designated as VLOPs/VLOSEs) and encourage the development of industry-wide best practices that protect user-rights more robustly.

Independent audits of this nature, if suitably adopted in Global South jurisdictions, can deliver similar transparency gains for all relevant stakeholders in the region. Some of the most egregious harms arising from interactions on platforms have been experienced the most acutely in the Global South.[28] Thus, stakeholders in the region have perhaps the most to benefit from the information that audits (contextualised to the relevant jurisdiction) would reveal.

At the same time, to derive meaningful insights from audits, jurisdictional regulators must possess the requisite regulatory capacity to process the information revealed through them. Requiring intermediaries to pay for audits, as laid down under the DSA, would preclude the primary financial burden of commissioning or conducting audits from falling on states. Nonetheless, regulatory bodies would require significant administrative and technical resources to study audit reports, verify their contents and draw learnings that can inform the evolution of regulation. Under the DSA, the analysis of audit reports to support the EC and the DSCs is a key statutory function of the European Board of Digital Services (EBDS).[29] Global South jurisdictions must address the preliminary challenge of mobilising adequate resources, if they are to institute audits of comparable scope as audits under the DSA. If imposed without equipping such bodies with the resources necessary for effective oversight,[30] audits may simply become a bureaucratic compliance-exercise prone to

27 Amelie P. Heldt, 'EU Digital Services Act: The White Hope of Intermediary Regulation' (2022), <https://library.oapen.org/bitstream/handle/20.500.12657/56979/1/978-3-030-95220-4.pdf#page=82> accessed 30 May 2024.

28 Anupam Chander, 'When the Digital Services Act Goes Global' (2023), <https://scholarship.law.georgetown.edu/facpub/2548/;> accessed 30 May 2024; Zahra Takhshid, 'Regulating Social Media in the Global South', 24 Vanderbilt Journal of Entertainment and Technology Law 1 (2022), <https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1564&context=jetlaw> accessed 30 May 2024.

29 DSA 2022, art 63(1)(b). *See also* DSA 2022, art 35(3).

30 Towards equipping regulatory bodies with financial resources for such supervision, the DSA imposes annual supervisory fees on VLOPs and VLOSEs. *See* DSA, art 43.

'street-lighting' – where auditors only examine those aspects of platforms on which platforms themselves cast light.[31] Audits of this nature would contribute little to meaningful transparency, while also draining time and resources away from potentially more impactful regulatory mechanisms.

Further, to maximise transparency gains for non-state stakeholders, it is critical that findings of audits are reported in sufficient detail to the general public. Public versions of audit reports must not be redacted excessively on the grounds such as confidentiality, avoidance of user-harm, security of intermediaries' services and public security. Accordingly, any such redactions must be scrutinised strictly so that they are proportional to the legitimate interest sought to be protected.[32] In the EU, the European Data Protection Supervisor and Data Protection Authorities set up under the GDPR, in view of their institutional mandates and expertise on matters relating to cybersecurity and data protection, can be crucial in assisting DSCs to harmonise these considerations. However, as we discussed in Chapter 1 (*Transparency in Recommendations*), many Global South jurisdictions have not yet passed data protection laws. Further, even where such laws have been passed, authorities overseeing their implementation are constrained severely in terms of their financial and technical capacities. Thus, in the event that audit reports are sought to be made public, there are concerns surrounding the principles on which transparency will be balanced with the interests of privacy and cybersecurity, as well as the suitable institutions to perform this exercise. Even in the EU, noting the dangers of the provision allowing for redaction under the DSA, certain stakeholders have called for making unredacted audit reports available to vetted researchers.[33]

31 Pershan (n 14).

32 AlgorithmWatch and AI Forensics, 'The DSA's Delegated Acts should strengthen a diverse auditing ecosystem for algorithmic risks' (2023), <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13626-Digital-Services-Act-conducting-independent-audits/F3424070_en> accessed 30 May 2024.

33 Algorithm Watch and AI Forensics, 'A diverse auditing ecosystem is needed to uncover algorithmic risks' (2023), <https://algorithmwatch.org/en/diverse-auditing-ecosystem-for-algorithmic-risks/> accessed 30 May 2024.

4.3. Auditing Criteria and Methodologies

A central element of any audit mechanism is a set of pre-defined criteria, against which the audited entity's compliance is assessed. Auditors under the DSA are meant to assess compliance of the VLOP/VLOSE with obligations specified in Chapter III, and any self-regulatory codes of conduct[34] and/or voluntary crisis protocols[35] drawn up under the DSA. Notably, determining compliance with many of these obligations would require significant exercise of discretion.[36] Article 39, for instance, requires each VLOP/VLOSE to make "reasonable efforts" to ensure that the information contained in its repository of advertisements is accurate and complete. Similarly, Article 35 requires them to institute "reasonable… mitigation measures", to address systemic risks arising from their services. To opine on whether a VLOP/VLOSE has complied with such obligations, auditors must, in effect, determine whether such efforts and measures pass the test of reasonableness. Moreover, they must inevitably consider the "geographical and/or social context" in which a systemic risk arises, and the specific groups that are affected.[37]

Such contextual and qualitative determinations are profoundly different from those required from audits in other sectors.[38] In the financial sector, for instance, auditors determine enterprises' compliance with reference to concretely defined accounting criteria. Even in environmental audits, certain quantitative metrics, such as carbon and water footprints, assist auditors in assessing enterprises' environmental impact

34 See DSA, art. 45.

35 See DSA, art. 45.

36 See Giovani De Gregorio and Oreste Pollicino, Auditing Platforms under the Digital Services Act (Verfassungsblog, 3 September 2024) <https://verfassungsblog.de/dsa-auditors-content-moderation-platform-regulation/#:~:text=According%20to%20the%20DSA%2C%20providers,certain%20codes%20of%20conduct%20and> accessed 22 January 2025.

37 AI Now Institute, 'Algorithmic Accountability: Moving Beyond Audits' (2023) <https://ainowinstitute.org/publication/algorithmic-accountability> accessed 19 August 2024.

38 Francisco Brito Cruz, Iná Jost and Catharina Vilela, 'In the second interview of the series, Tom Barraclough talks about auditing mechanisms for platforms', InternetLab (2023), <https://internetlab.org.br/en/news/in-the-second-interview-of-the-series-tom-barraclough-talks-about-auditing-mechanisms-for-platforms/> accessed 30 May 2024.

with some objectivity. In the absence of similarly objective criteria to audit compliance with the DSA's obligations, many stakeholders have expressed apprehension regarding auditors' ability to provide reliable and justifiable opinions.[39]

Towards providing more clarity on auditing criteria and procedures, the EC adopted the delegated legislation on audits under the DSA (Delegated Regulation') in October 2023. In effect, the Delegated Regulation leaves the task of formulating suitable auditing criteria on the audited entity and the auditors. At the first instance, the VLOP/VLOSE must disclose to auditors its internally–formulated benchmarks towards monitoring compliance with the DSA.[40] These benchmarks then form the basis on which auditors formulate the auditing criteria.[41] Similarly, while the Delegated Regulation provides a broad procedural framework for audits, it leaves the precise methodology to the determination of auditors, on a case-to-case basis.[42] Before the audit commences, auditors must analyse the 'audit risks', i.e. the risks that they express an incorrect audit opinion.[43] They must accordingly design appropriate methodologies for conducting the audits, so as to minimise such audit risks to a level where they can express their final audit opinions "at a reasonable level of assurance".[44]

As scholars have noted, platform-audits, and more broadly, human rights audits, are an emergent accountability mechanism, even in the Global North.[45] As a result, the

39 Francisco Brito Cruz and others (n 38); Jason Peilemeier, Ramsha Jahangir and Hillary Ross, 'Ensuring Digital Services Act Audits Deliver on Their Promise' Tech Policy Press (19 February, 2023) <https://www.techpolicy.press/ensuring-digital-services-act-audits-deliver-on-their-promise/> accessed 19 August 2024 .

40 Delegated Regulation, art 5(1)(a).

41 Delegated Regulation, art 10(2)(a)

42 Delegated Regulation, art section IV.

43 Delegated Regulation, art 9.

44 Delegated Regulation, art 10(1).

45 Vogus and Lanso (n 1).

benchmarks and methodologies for such audits are in early stages of development.[46] Against this backdrop, the flexibility granted to auditors in developing auditing criteria and procedures can be understood as a pragmatic legislative choice. It allows auditors to devise and tailor their audits to the distinctive service(s) each VLOP/VLOSE offers, the distinctive system(s) that it employs, and the distinctive risk(s) that it poses in a particular jurisdiction. It also allows auditors to modify the auditing criteria and procedures, albeit within a broad legislative framework, on the basis of evidence gathered in the course of an audit.[47] Importantly, it leaves room for iterative development of criteria and procedures through successive auditing cycles, in two significant ways – first, on the basis of insights gained from previous audit reports; [48] and second, through guidance from the EC, the EBDS and other authoritative sources, as well as through case-law, enforcement-related decisions, vetted researchers and public consultations under the DSA.[49]

Setting out granular auditing criteria and procedures in law may indeed be impractical, and even undesirable, particularly in these initial years of platform-audits. At the same time, the value of the audit as an accountability mechanism rests on the comparability of audits across entities and across auditing cycles. As commentators have observed, auditing standards and methodologies are central to the comparability and legitimacy of audits.[50] If auditors employ widely divergent auditing standards, or follow disparate methodologies, it would be impossible to draw any reliable ecosystem-wide conclusions from audit processes and reports. The

46 See, for instance, Anna-Katharina Meßmer and Martin Degeling, 'Auditing Recommender Systems' Stiftung Neue Verantwortung (2023), where a methodological framework is proposed for auditing recommender systems under the DSA <<https://arxiv.org/ftp/arxiv/papers/2302/2302.04556.pdf> accessed 30 May 2024.

47 *See* Delegated Regulation art 10, which allows auditors to modify the auditing criteria as well the auditing procedure in the course of an audit.

48 *See*, for instance, Delegated Regulation art 8(1)(b)(i), which authorises auditors to comment on the appropriateness of compliance benchmarks employed by the VLOP, regardless of their opinion on the VLOP's compliance against such benchmarks.

49 See Delegated Regulation, recital 16.

50 Ellen P. Goodman and Julia Trehu, 'AI Audit-Washing and Accountability' GMF (2022), <https://www.gmfus.org/news/ai-audit-washing-and-accountability> accessed 30 May 2024; Jason Peilemeier and others (n 39).

AI Now Institute has highlighted how without clear benchmarks, broadly-scoped algorithmic audits are already being used by powerful platforms to preclude more substantive and contextual inquiries on their business models.[51] Self-adopted auditing benchmarks and methodologies could heighten the risk of such 'audit-washing', where audits are exploited by platforms to legitimise their practices and evade accountability.[52] Separately, lack of clarity around auditing standards can invoke disproportionate fear of statutory liability or reputational harm amongst auditors,[53] undermining the performance of effective audits and the development of a competitive auditing ecosystem.

With the introduction of platform-audits and impact assessments in Global North legislation, initiatives for the development of auditing standards are set to gather momentum. Considering the partly-technical character of platform-audits, established standardisation bodies working on technical standards could take up the task of formulating auditing standards and procedures. The DSA, in fact, expressly envisages promoting the development of audit-related standards at European and international standardisation bodies.[54] It remains to be seen how standardisation initiatives will engage with the aforementioned differences in the nature of and the risks posed by platforms' diverse services.

Given the pre-eminence of international standardisation bodies, auditing standards and methodologies developed through their initiatives can ossify to become the

51 AI Now Institute (n 37).

52 Anna-Katharina Meßmer and Martin Degeling, 'Auditing Recommender Systems' Stiftung Neue Verantwortung (2023) <https://arxiv.org/ftp/arxiv/papers/2302/2302.04556.pdf> accessed 30 May 2024; Ellen P. Goodman and Julia Trehu, 'AI Audit-Washing and Accountability' GMF (2022), <https://www.gmfus.org/news/ai-audit-washing-and-accountability> accessed 30 May 2024; Algorithm Watch and AI Forensics, 'A diverse auditing ecosystem is needed to uncover algorithmic risks' (2023), <https://algorithmwatch.org/en/diverse-auditing-ecosystem-for-algorithmic-risks/> accessed 30 May 2024; Sebastian Klovig Skelton, 'AI accountability held back by 'audit-washing' practices' (Computer Weekly, 23 November 2022) <https://www.computerweekly.com/news/252527612/AI-accountability-held-back-by-audit-washing-practices> accessed 19 August 2024.

53 Global Network Initiative, 'GNI Submission on the DSA Delegated Regulation on Independent Audits' (2023) <https://globalnetworkinitiative.org/wp-content/uploads/2023/06/GNI-DSA-Audits-Comments.pdf> accessed 30 May 2024; Goodman and Trehu (n 50).

54 DSA 2022, art 44(1)(e).

"ceilings for performance".[55] Moreover, there are structural obstacles to integrating human rights considerations in standards-development initiatives, especially at such bodies. Prominent standardisation bodies are typically dominated by engineers and other members of the technical community.[56] Even where civil society organisations and human rights experts gain access to such forums, their effective participation is hindered by the highly technical language used in discussions, as well as the costs of participation.[57] As highlighted in a response to a recent consultation on technical standards and human rights led by the UN High Commissioner for Human Rights, these challenges disproportionately affect participation from the Global South.[58] Most international standardisation bodies and their respective working groups convene in the Global North. Further, civil society organisations in the Global South have fewer resources than their Global North counterparts and rely on support from their government or funders from the Global North. This severely constrains their ability to participate meaningfully and submit their independent inputs to international standardisation bodies.

For Global South states contemplating a framework for platform-audits, facilitating meaningful participation of domestic experts in such initiatives would be crucial, so that such standards and methodologies adequately account for their distinctive contexts and interests. Towards this end, such states should steadily build domestic capacity to influence these initiatives. As a starting point, they should consider placing platform-audits on the agendas of their national standards bodies, and

55 Goodman and Trehu (n 50).

56 UN OHCHR, 'OHCHR consultation on human rights and technical standard-setting processes for new and emerging digital technologies' (2023) <https://www.ohchr.org/en/events/events/2023/ohchr-consultation-human-rights-and-technical-standard-setting> accessed 30 May 2024.

57 ibid.

58 Data Privacy Brasil Research Association, 'Submission by Data Privacy Brasil Research Association to the call for inputs: "The relationship between human rights and technical standard-setting processes for new and emerging digital technologies (2023)" - Report of the High Commissioner for Human Rights' (2023) <https://www.ohchr.org/sites/default/files/documents/issues/digitalage/cfis/tech-standards/subm-standard-setting-digital-space-new-technologies-csos-data-privacy-brazil-research-association-3-input-part-2.pdf> accessed 30 May 2024.

constitute focus-groups to deliberate on suitable standards and methodologies. Such focus-groups must include, and facilitate active engagement with, human rights practitioners, civil society organisations and researchers on platform governance. On one hand, this would guard against the hijacking of such processes by dominant platforms; concomitantly, it would ensure that human rights considerations remain central to the evolution of auditing standards.

As the first set of audit reports under the DSA is published towards the end of 2024, Global South states must facilitate rigorous assessments of these reports, focussing on the auditing standards and methodologies disclosed therein. Given substantial user-bases in Global South states, many intermediaries designated as VLOPs/VLOSEs under the DSA are also likely to be the subjects of any prospective platform-regulation in Global South states. Thus, stakeholders in such states should give due consideration to whether the standards and methodologies developed in DSA-audits, reveal meaningful information regarding these intermediaries' systems and can be suitably adapted to align with divergent contexts in the Global South. This will be particularly useful for states introducing or proposing to introduce substantive obligations similar to those under the DSA.

4.4. Auditor Selection

The audit mechanism under the DSA derives much of its objectivity and resultant legitimacy from requirements relating to auditors' expertise and independence.

Expertise

As discussed above, auditors under the DSA are required to possess demonstrable expertise in risk management. Further, they must demonstrate adequate technical competence and capabilities, including systems capable of maintaining necessary levels of confidentiality.

Considering the novelty and the extensive scope of platform-audits, significant cross-disciplinary expertise will be required to effectively conduct an audit in a time-bound manner. The implementation of audits under the DSA is expected to result in a

supply of audit-related services over time.[59] However, there are serious concerns surrounding the present capacity to conduct such audits, even in the Global North.[60] According to an estimate from 2021, only around 10 to 20 reputable firms across the world offered algorithmic auditing services.[61] Such capacity concerns are likely to be particularly acute in Global South nations, where stakeholders (including researchers and other technical experts) have hitherto had little direct access to information on platforms' functioning and scarce resources for research. Thus, Global South jurisdictions attempting to adopt audit-mechanisms must formulate realistic criteria for the selection of auditors, accounting for the present scarcity of organisations capable of conducting audits. Such criteria can be made more stringent progressively, as a competitive supply of auditors emerges over successive audit cycles.

At least in the initial years, platform-audits may have to be conducted collaboratively by consortiums of organisations and experts pooling in their resources. In fact, the Delegated Regulation expressly allows for the engagement of multiple organisations for conducting DSA audits, both jointly and as sub-contractors.[62] Such consortiums are likely to include established business-consulting and audit firms (providing audit services across domains, such as the "Big Four"), organisations advising businesses on risk management, corporate social responsibility and information security, as well as other technical experts.[63] Such actors may have the financial and technical resources required to conduct platforms-audits. However, they may have limited experience with examining issues relating to human rights and platform accountability, which civil society groups and researchers have developed over recent

59 Claire Pershan, 'As the Digital Services Act takes shape, are platform accountability experts at a crossroads?' (Mozilla, 2023), <https://foundation.mozilla.org/en/blog/digital-services-act-and-platform-accountability/> accessed 30 May 2024.

60 Jason Peilemeier and others (n 39).

61 Alfred Ng, 'Can Auditing Eliminate Bias from Algorithms?' (The MarkUp, 2021) <https://themarkup.org/the-breakdown/2021/02/23/can-auditing-eliminate-bias-from-algorithms> accessed 30 May 2024.

62 Delegated Regulation, recital 3.

63 In fact, the Delegated Regulation expressly allows the way for engagement of multiple entities as auditors.

years.[64] Thus, it is concerning that the DSA does not require auditors to possess expertise in human rights impact assessments.

Independence

To protect auditors' independence from the VLOP/VLOSE they audit, the DSA requires that they have proven objectivity and professional ethics, based on relevant codes of practice or standards. Crucially, they must not have any conflict of interest with the VLOP/VLOSE, or any connected legal person.

As evidence from the financial sector affirms, the quality of audits tends to suffer significantly if the audited entity has engaged or expresses willingness to engage the auditor for non-audit services.[65]Accordingly, the DSA requires that auditors must not have provided the VLOP/VLOSE (or any connected legal person) any "non-audit services related to the matters audited", in the 12 months prior to the commencement of the audit.[66] Further, they must commit not to provide them any non-audit service for a period of 12 months after its completion.[67] They also must not have provided the same VLOP/VLOSE (or any connected legal person) audit services under the DSA, for a period of 10 consecutive years or more.[68] Lastly, while VLOPs and VLOSEs are required to commission the audits, the fees payable to the auditors must not be contingent on the audit's results.[69] Such criteria are designed to ensure that unlike internal audits, which are susceptible to bias by their very nature,[70] audits under the DSA are carried out by external functionaries in an unbiased manner.

64Pershan (n 59).

65 Monika Causholli and others, 'Future Nonaudit Service Fees and Audit Quality' (2014), <https://gattonweb.uky.edu/FACULTY/PAYNE/acc490/Graduate%20Student%20Articles/CAR%20Final.pdf> accessed 30 May 2024.

66 DSA 2022, art 37(3)(a)(i).

67 DSA 2022, art 37(3)(a)(i).

68 DSA 2022, art 37(3)(a)(ii).

69 DSA 2022, art 37(3)(a)(iii).

70 Vogus and Lanso (n 1).

Despite these criteria for independence, Laux *et al* have highlighted that, audits under the DSA, financed by VLOPs/VLOSEs themselves, are at risk of 'capture' by them.[71] Drawing lessons from accounting audits, Laux *et al* predict that due to the concentrated structure of audit markets (reinforced by the VLOP/VLOSE threshold), VLOPs/VLOSEs will dominate the demand-side. Consequently, driven by standard economic incentives, auditors will tend to cater to the interests of VLOPs/VLOSEs and draw up favourable audit opinions to continue to receive auditing assignments. The potential for such audit-capture can be particularly high in Global South jurisdictions, where at present, only a handful of consulting firms may have the resources to conduct audits of scope comparable to the DSA. Such firms' positions as advisors or service-providers to the same VLOPs/VLOSEs in other domains, such as accounting and financial risk- management, could also constrain their effective independence as auditors.

In these circumstances, it is crucial that actors accountable to the broader public yet sufficiently independent from the state, and with experience in examining concerns relating to human rights, engage extensively with auditing processes. As Raji *et al* have observed, the efforts of civil society actors, investigative journalists, lawyers and other third parties have been pivotal in unearthing numerous instances of harm caused by such systems.[72] Without the engagement of such actors, platform-auditing could devolve into a purely technical exercise that fails to account for the socio-technical character of platforms.[73] Such engagement can be direct, by the inclusion of such actors in auditing consortiums, standardisation initiatives and voluntary auditing procedures under collaborations like the Global Network Initiative.

71 Johann Laux, Sandra Wachter and Brett Mittelstadt, 'Taming the Few: Platform Regulation, Independent Audits, and the Risks of Capture Created by the DMA and the DSA', Computer Law & Security Review 43 (2021): 105613, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4096655> accessed 30 May 2024.

72 'Inioluwa Deborah Raji and others, 'Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance' In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (2022) <https://www.skillscommons.org/bitstream/handle/taaccct/18870/Raji_et_al_2022_Outsider_Oversight.pdf?sequence=3&isAllowed=y> accessed 30 May 2024.

73 Jason Peilemeier and others (n 39).

Additionally, such engagement can take the form of knowledge-sharing between technical and non-technical functionaries at multi-stakeholder forums, including existing ones such as the Internet Governance Forum and RightsCon. To enhance such engagement, some have also advocated for direct regulatory pathways for inputs from public interest researchers, including those not "vetted" under the DSA, to inform auditing processes – for example, by requiring auditors to consider such inputs as audit-evidence.[74]

Noting independence-related concerns embedded in audits commissioned by audited entities, AlgorithmWatch and AI Forensics argue for the recognition of adversarial audits in law.[75] Unlike second-party audits by auditors commissioned by the audited entity, these would be conducted by independent third parties having no contractual relationship with the audited entity – such as civil society watchdogs, algorithmic accountability experts, and other human rights practitioners. If Global South jurisdictions were to consider instituting third-party audits, the primary challenge would be identifying truly independent third-parties in such states. In the *status quo*, civil society organisations in the region are severely restricted in their capacity to mobilise resources. Often faced with uncertain research environments, they rely on resources from platforms themselves, or from other funders with potential conflicts of interest. Further, the heterogeneity of civil society, and the diversity of interests therein, would represent another major challenge in identifying third-parties that can legitimately claim to represent the public in such states.

74 Sasha Costanza-Chock, 'Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem' ACM Digital Library (2022) <https://arxiv.org/pdf/2310.02521.pdf.> accessed 30 May 2024.

75 AlgorithmWatch and AI Forensics (n 52); Sasha Costanza-Chock (n 74).

Insights for the Global South

- Regulatory audits, conducted by external auditors and overseen by regulatory authorities, can be an effective mechanism to systematically illuminate platforms' systems, policies and procedures. The information gathered through such audits can potentially assist stakeholders in understanding the propagation of information via platforms in the Global South, and in affixing accountability on platforms for the adverse effects of their services in the region.

- Global South states must equip relevant regulatory bodies with adequate resources and independent powers to meaningfully process and critically assess audit reports under platform-audit frameworks, verify their contents and draw learnings that can inform the evolution of platform regulation.

- Clear benchmarks and methodologies are central to the reliability of audits, without which audits can be exploited by platforms to evade accountability. At the same time, auditing procedures must respect differences between platforms and the risks they pose in divergent contexts. Accordingly, Global South states should formulate benchmarks and methodologies tailored to their respective jurisdictional contexts as well as to differences between the risks posed by different kinds of services. As an initial step, they should build capacity to formulate such benchmarks and methodologies, and to contribute meaningfully in international initiatives, including multistakeholder forums and standard-setting bodies.

- Only very few organisations across the world currently possess the resources and expertise to conduct audits. Such limitations are particularly acute in the Global South. This heightens the risk of audit-capture by platforms, particularly if audits are commissioned by platforms themselves. Accordingly, Global South states should consider fostering an ecosystem of independent audits conducted by third parties acting in the public interest.

- In any case, given that issues relating to human rights and platform accountability have been extensively and predominantly examined by civil society organisations, independent researchers and other human rights practitioners, auditing frameworks in

the Global South must provide pathways for the active engagement of such third parties in auditing as well the processes for formulation of auditing benchmarks and methodologies.

- Global South states must navigate limitations on their regulatory capacity and equip regulatory bodies with adequate financial and technical resources, and independent powers to meaningfully assess audit reports and draw learnings that can inform the evolution of platform regulation.

- Towards maximising the transparency gains from audits, audit reports must be made public. While it may be necessary to redact certain information from audit reports, any redaction must be strictly proportional to the countervailing interest sought to be protected. Global South states should institute robust data protection legislation, to meaningfully balance transparency alongside privacy considerations.

5. RESEARCHER ACCESS TO PLATFORM DATA

5.1. Introduction

As concerns about misinformation, hate speech, extremist content, and online safety grow, it has increasingly become clear that regulators, academia and civil society have insufficient information to gauge the magnitude, spread and impact of these harms.[1] The recent series of whistleblower revelations have only served to highlight the glaring information asymmetry that exists between platforms and other stakeholders.[2] As a result, external observers cannot understand how platform design choices, internal processes and algorithms contribute to online harms. In this context, data access for public interest research becomes an important mechanism to hold platforms accountable.[3] Such research can shed light on how existing content

1 See Robert Gorwa and Timothy Garton Ash, 'Democratic Transparency in the Platform Society', *Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge University Press 2020) <https://www.cambridge.org/core/books/social-media-and-democracy/democratic-transparency-in-the-platform-society/F4BC23D2109293FB4A8A6196F66D3E41> accessed 10 November 2023; Axel Bruns, 'After the "APIcalypse": Social Media Platforms and Their Fight against Critical Scholarly Research' (2019) 22 Information, Communication & Society 1544 <https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1637447> accessed 11 April 2024; Ganaele Langlois and Greg Elmer, 'The Research Politics of Social Media Platforms' (2013) 14 Culture machine.

2 Georgia Wells, Jeff Horwitz and Deepa Seetharaman, 'Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show' *Wall Street Journal* (14 September 2021) <https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739> accessed 11 July 2023; Cristiano Lima, 'A Whistleblower's Power: Key Takeaways from the Facebook Papers' *The Washington Post* (25 October 2021) <https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/>.

3 Mathias Vermeulen, 'The Keys to the Kingdom' (*Knight First Amendment Institute at Columbia University*) <http://knightcolumbia.org/content/the-keys-to-the-kingdom>; Philipp Darius and Daniela Stockmann, 'Implementing Data Access of the Digital Services Act' <https://opus4.kobv.de/opus4-hsog/frontdoor/deliver/index/docId/4947/file/Implementing_Data_Access_Darius_Stockmann_2023.pdf> accessed 14 February 2024; Paddy Leerssen, Amélie Heldt and Matthias C Kettemann, 'Scraping By?: Europe's Law and Policy on Social Media Research Access' <https://www.ssoar.info/ssoar/handle/document/86427> accessed 8 February 2024; Nathaniel Persily, 'A Proposal for Researcher Access to Platform Data: The Platform Transparency and

moderation, recommender systems, and advertising models impact society. Researchers can act as "knowledge creators, policy advisors, policy watchdogs and social innovators" driving new regulation, platform policies and design.[4] Data access also provides an invaluable resource for social science researchers to study the evolving social, political and cultural discourses and their impact on individuals, communities and democracies.[5]

Online hate speech and misinformation can prove to be catastrophic, especially in regions of conflict, countries under authoritarian regimes and jurisdictions with weak rule of law. Here, independent public interest research becomes extremely vital to not only ensure platform accountability but also to safeguard the human rights of the most vulnerable segments of the population. This holds true in several Global South countries, where independent investigations have revealed that platforms like Facebook have played a role in facilitating violence against religious and ethnic minorities.[6]

Accountability Act' (2021) 1 Journal of Online Trust and Safety <https://tsjournal.org/index.php/jots/article/view/22> accessed 19 January 2023.

4 Darius and Stockmann (n 3).

5 Homero Gil de Zúñiga and Trevor Diehl, 'Citizenship, Social Media, and Big Data: Current and Future Research in the Social Sciences' (2017) 35 Social Science Computer Review 3 <https://doi.org/10.1177/0894439315619589> accessed 12 May 2024; Pablo Barberá, 'Social Media, Echo Chambers, and Political Polarization' [2020] Social Media and Democracy: The State of the Field, Prospects for Reform 34; Bruns (n 1).

6 Take Facebook's role in creating an echo chamber of hatred leading to the dehumanisation of Rohingya Muslims in Myanmar, culminating in ethnic cleansing in 2017, which was established by several independent fact-finding bodies, including the UN fact-finding mission. Similarly, an independent New York Times investigation revealed how Facebook fanned violence in the 2018 Lybian war, with military-grade weapons being openly traded on the platform and armed groups creating their own Facebook pages. More recently, independent investigations have uncovered how inflammatory content on Facebook is contributing to ethnic massacres in war-torn Ethiopia. In India, too, whistleblower leaks have confirmed what several independent journalists have long suspected: Facebook is awash with hate speech and conspiracy theories despite such content being internally flagged by employees. See Amnesty International, 'The Social Atrocity: Meta and the Right to Remedy for the Rohingya' (2022) <https://www.amnesty.org/en/documents/asa16/5933/2022/en/>; Reuters, 'Myanmar: UN Blames Facebook for Spreading Hatred of Rohingya' *The Guardian* (13 March 2018) <https://www.theguardian.com/technology/2018/mar/13/myanmar-un-blames-facebook-for-spreading-hatred-of-rohingya> accessed 5 June 2023; Declan Walsh and Suliman Ali Zway, 'A Facebook War: Libyans Battle on the Streets and on Screens' *The New York Times* (4 September 2018) <https://www.nytimes.com/2018/09/04/world/middleeast/libya-facebook.html> accessed 21

5.2. Existing Voluntary Mechanisms of Data Access are Insufficient and Precarious

Until now, research on platforms has typically relied on access to datasets through public Application Programming Interfaces (APIs) and other data-sharing tools made available voluntarily by platforms,[7] data-sharing arrangements between research organisations and platforms,[8] and independent data collection methods like scraping of public data by researchers,[9] and data donation by users.[10] This has meant that researchers have almost always been dependent on the platform's discretion and goodwill to gain access to data.[11]

Those who accessed data through platform APIs were subject to their terms and conditions, while the platforms assumed little accountability for the quality and veracity of the data they provided.[12] Researchers are bound by the terms set by

December 2021; Jasper Jackson, Mark Townsend and Lucy Kassa, 'Facebook "Lets Vigilantes in Ethiopia Incite Ethnic Killing"' *The Observer* (20 February 2022) <https://www.theguardian.com/technology/2022/feb/20/facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing> accessed 5 June 2023; Billy Perrigo, 'Facebook Let an Islamophobic Conspiracy Theory Flourish in India Despite Employees' Warnings' [2021] *Time*<https://time.com/6112549/facebook-india-islamophobia-love-jihad/>.

7 Langlois and Elmer (n 1).

8 'SOCIAL SCIENCE ONE' <https://socialscience.one/home> accessed 11 April 2024.

9 See Leerssen, Heldt and Kettemann (n 3).

10 See Irene I van Driel and others, 'Promises and Pitfalls of Social Media Data Donations' (2022) 16 Communication Methods and Measures 266 <https://doi.org/10.1080/19312458.2022.2109608> accessed 12 May 2024.

11 See Catherine Altobelli and others, 'To Scrape or Not to Scrape? The Lawfulness of Social Media Crawling under the GDPR' (2021); 'Status Report: Mechanisms for Researcher Access to Online Platform Data' (EC 2024) <https://digital-strategy.ec.europa.eu/en/library/status-report-mechanisms-researcher-access-online-platform-data>; Megan A. Brown, Josephine Lukito, and Kai-Cheng, 'What Does CrowdTangle's Demise Signal for Data Access Under the DSA?' (*Tech Policy Press*, 27 March 2024) <https://techpolicy.press/what-does-crowdtangles-demise-signal-for-data-access-under-the-dsa> accessed 4 April 2024; Brandi Geurkink Gilbert Sarah, 'Why Reddit's Decision to Cut off Researchers Is Bad for Its Business—and Humanity' (*Fast Company*, 22 January 2024) <https://www.fastcompany.com/91014116/reddit-researchers-bad-for-business> accessed 2 September 2024.

12 Cynthia O'Murchu, Jemima Kelly and David Blood, 'Facebook under Fire as Political Ads Vanish from Archive' *Financial Times* (10 December 2019) <https://www.ft.com/content/e6fb805e-1b78-11ea-97df-cc63de1d73f4> accessed 3 May 2023; Leerssen, Heldt and Kettemann (n 3).

platforms on rate limits and quotas while accessing their APIs.[13] This frequently left researchers at the mercy of platform management, as was painfully evident when Facebook restricted access to its public API in the aftermath of the Cambridge Analytica scandal.[14] More recently, Twitter abruptly discontinued free access to its API and made the revised rates unaffordable for independent researchers and a majority of research organisations, especially in the Global South.[15] Twitter also substantially increased the price for another research access API called "Decahose" and demanded that researchers delete all data unless they pay the new unaffordable rates.[16] Similarly, the recently introduced Twitter Moderation Research Consortium (TMRC) has been stalled after an employee exodus[17] and Meta has systematically dismantled and retired CrowdTangle,[18] which was widely used by researchers and journalists to study the spread of misinformation and hate speech on its platforms.[19] Not only have these mechanisms been precarious, but voluntary measures have also

13 'Status Report: Mechanisms for Researcher Access to Online Platform Data' (n 11).

14 Ben Smee, 'Facebook's Data Changes Will Hamper Research and Oversight, Academics Warn' *The Guardian* (25 April 2018) <https://www.theguardian.com/technology/2018/apr/25/facebooks-data-changes-will-hamper-research-and-oversight-academics-warn> accessed 3 June 2023.

15 Ivan Mehta, 'Twitter's Restrictive API May Leave Researchers out in the Cold' (*TechCrunch*, 14 February 2023) <https://techcrunch.com/2023/02/14/twitters-restrictive-api-may-leave-researchers-out-in-the-cold/> accessed 3 June 2023.

16 Chris Stokel-Walker, 'Twitter Is Making Researchers Delete Data It Gave Them Unless They Pay $42,000' (*inews.co.uk*, 25 May 2023) <https://inews.co.uk/news/twitter-researchers-delete-data-unless-pay-2364535> accessed 29 May 2023.

17 Sheila Dang, 'Twitter Research Group Stall Complicates Compliance with New EU Law' (*euronews*, 28 January 2023) <https://www.euronews.com/next/2023/01/28/twitter-moderation-insight>.

18 Rebecca Bellan, 'Meta Axed CrowdTangle, a Tool for Tracking Disinformation. Critics Claim Its Replacement Has Just "1% of the Features"' (*TechCrunch*, 15 August 2024) <https://techcrunch.com/2024/08/15/meta-shut-down-crowdtangle-a-tool-for-tracking-disinformation-heres-how-its-replacement-compares/> accessed 2 September 2024; Megan A. Brown, Josephine Lukito, and Kai-Cheng (n 11); Sarah Scire, 'A Window into Facebook Closes as Meta Sets a Date to Shut down CrowdTangle' (*Nieman Lab*, 14 March 2024) <https://www.niemanlab.org/2024/03/a-window-into-facebook-closes-as-meta-sets-a-date-to-shut-down-crowdtangle/> accessed 28 May 2024.

19 John Albert, 'Facebook's Gutting of CrowdTangle: A Step Backward for Platform Transparency' (*AlgorithmWatch*, 3 August 2022) <https://algorithmwatch.org/en/crowdtangle-platform-transparency/> accessed 3 June 2023.

provided platforms with the power to vet researchers and research projects raising substantial conflicts of interest.[20]

The researchers who gained access through data-sharing agreements like the Social Science One spearheaded by Harvard University[21] also ran into problems without proper accountability mechanisms, as platforms provided far less data than was originally promised and also provided incorrect data,[22] leading to substantial delays in ongoing research projects. Without regulatory backing and independent audit mechanisms, it is very hard to detect both genuine errors and malicious tampering of data provided by platforms. Moreover, these arrangements require cooperation between institutes and platforms, which is highly dependent on the resources and the negotiating power of such institutions. Global South institutions and researchers typically have less experience and limited power to negotiate and enforce terms with platforms.[23]

Researchers employing independent data collection methods like data scraping have also run into troubles with platforms that often sue researchers for violating their Terms of Service.[24]

---

20 Meta has recently, partnered with the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan application review process for its Content Library through to review applications to its Content Library. 'Meta Content Library and API | Transparency Centre' <https://transparency.meta.com/en-gb/researchtools/meta-content-library/> accessed 29 May 2024.

21 'SOCIAL SCIENCE ONE' (n 8).

22 Craig Timberg, 'Facebook Made Big Mistake in Data It Provided to Researchers, Undermining Academic Work' *Washington Post* (11 September 2021) <https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/> accessed 5 June 2023.

23 See Jhalak Kakkar, 'Tackling Misinformation in Emerging Economies and the Global South: Exploring Approaches for the Indian Context', *Digital Technologies in Emerging Countries* (Stanford Cyber Policy Center 2023) <https://cddrl.fsi.stanford.edu/news/digital-technologies-emerging-countries>.

24 See Jeff Horwitz, 'WSJ News Exclusive | Facebook Seeks Shutdown of NYU Research Project Into Political Ad Targeting' *Wall Street Journal* (23 October 2020) <https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533> accessed 29 May 2023.

Even when granting data access to researchers was an important component of the EU Code of Practice on Disinformation,[25] it did not result in any meaningful access to granular platform data beyond the limited access through platform APIs and advertisement archives.[26]

5.3. Researcher Access to Platform Data in the DSA

As the importance of studying platform data becomes evident, regulators, especially in the EU[27] and the USA,[28] are looking at researcher access to platform data as an important transparency mechanism. In the Global South, Brazil's draft Internet Freedom, Responsibility and Transparency Law (PL 2630/2020) also contained provisions for researcher access to platform data.[29] While discussions across jurisdictions are at different stages, DSA has become the first legislation to mandate data access for research purposes with respect to very large online platforms (VLOPs) and very large online search engines (VLOSEs).

The DSA envisages researcher access to be complementary to other transparency mechanisms like risk assessment by platforms (see Chapter 3) and audits by third-party experts (see Chapter 4), as well as regulatory monitoring by Digital Service Coordinators (DSCs).[30] It lays down two mechanisms for granting researchers access

25 See the Code of Practice on Disinformation 2018 and The Strengthened Code of Practice on Disinformation 2022<https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation>.

26 Mathias Vermeulen (n 3).

27 In the EU, the Code of Practice on Disinformation provided for voluntary mechanisms for researcher access to platform data.

28 In the US, the Platform Accountability and Transparency Act, Social Media Data Act, Digital Services Oversight and Safety Act are some of the legislations under consideration. See Caitlin Vogus, 'Independent Researcher Access to Social Media Data: Comparing Legislative Proposals' (2022) <https://cdt.org/insights/independent-researcher-access-to-social-media-data-comparing-legislative-proposals/>.

29 The Internet Freedom, Responsibility and Transparency Law (2020) [2.630] <https://cyberbrics.info/wp-content/uploads/2021/06/Brazilian-Fake-News-Draft-Bill-no.-2.630-of-2020.pdf>.

30 DSA 2022, recital 96.

to platform data. The first[31] mandates VLOPs and VLOSEs to provide access to "vetted researchers" for conducting research that contributes to the "detection, identification and understanding of systemic risks in the Union"[32] and assessment of risk mitigation measures.[33] The second mandates VLOPs and VLOSEs to provide access to public data to researchers affiliated with not-for-profit bodies, organisations and associations for the detection, identification and understanding of systemic risks.[34]

While the DSA outlines the process of applying for research proposals,[35] baseline conditions for vetting researchers by the DSCs,[36] and grounds for platforms to seek amendment to data requests,[37] the operational and procedural details have been left to delegated legislation to be adopted by the European Commission (EC).[38]

31 DSA 2022, art 40(4).

32 These are laid out in art 34(1) of the DSA.

33 These are laid out in art 35 of the DSA.

34 DSA 2022, art 40(12).

35 DSA 2022, arts. 40(8) & (9).

36 DSA 2022, art 40(8).

37 DSA 2022, art 40(5).

38 As per article 40(13) of the DSA, the EC after consulting the Board, shall adopt delegated acts laying down the technical conditions under which VLOPs and VLOSEs have to share data. In order to lay down the specific conditions for data sharing that take into account the rights and interests of all stakeholders, the EC conducted a call for evidence in April-May 2023 gathering feedback from researchers, civil society organisations, online platforms and other stakeholders. After receiving feedback from stakeholders through public and expert-level consultations, the Commission has recently published the Draft Delegated Regulation laying down the "Technical Conditions and Procedures under Which Providers of Very Large Online Platforms and of Very Large Online Search Engines Are to Share Data Pursuant to Article 40 of Regulation" in October 2024 for public consultation. *See* European Commission, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (*European Commission - Have your say*) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act_en> accessed 16 January 2024.

5.4. Scope of Research

The DSA links the purpose of research to systemic risks in the EU. Vetted researchers under Article 40(4) of the DSA can conduct research contributing to the "detection, identification and understanding" of systemic risks as well as the "adequacy, efficiency and impact" of the risk mitigation measures taken by platforms. Meanwhile, researchers under Article 40(12) can use public data provided by platforms to contribute to the "detection, identification and understanding" of systemic risks.

As discussed in Chapter 3 (Risk Management), the systemic risks identified under the DSA include dissemination of illegal content, negative impact on exercising fundamental rights under the EU Charter, civic discourse, electoral processes, public security, gender-based violence, public health, minors and an individual's physical and mental well-being.[39] While these risks are defined broadly, it is important to note that the law does not leave the research scope completely open-ended.[40]

One reason for linking data access for researchers to systemic risks may be to provide independent oversight over the platform's own risk assessments. Insights from independent research can help corroborate, verify or challenge the findings presented in platforms' risk assessment reports. The DSA also envisages the studies conducted by vetted researchers to contribute to independent audits.[41] This makes data access to research an important mechanism in the enforcement of the DSA.[42]

Although the risk categories are defined sufficiently broadly, this limiting of research scope restricts academic freedom,[43] possibly rendering certain important research

39 DSA 2022, art 34(1).

40 Paddy Leerssen, 'Platform research access in Article 31 of the Digital Services Act: Sword without a shield?' [2021] Verfassungsblog<https://verfassungsblog.de/power-dsa-dma-14/> accessed 29 May 2023.

41 DSA 2022, recital 92.

42 See Leerssen (n 40); Mathias Vermeulen (n 3).

43 It is interesting to note that Article 13 of the EU Charter of Fundamental Rights, protects academic freedom: "The arts and scientific research shall be free of constraint. Academic freedom shall be

questions untenable under the DSA. Thus, certain scholars have argued for leaving the research scope to the more open-ended "public interest or scientific research".[44]

Empowering states to define research agendas must always be viewed with some caution. This becomes especially significant in some Global South countries, where research aligned to state interests might use data access to target dissenting speech.[45]

By linking the purpose of research to human rights under the Charter of Fundamental Rights of the EU, the DSA appears to mitigate against some of these risks associated with potential abuse. However, terms like "public security" and "civic discourse" can steer research in a direction that serves state interests, especially in authoritarian countries and can even become a threat to dissenting individuals and minority communities. This can be gauged from the fact that many states, like India,[46] Bangladesh,[47] Pakistan,[48] Nigeria,[49] Turkey,[50] Iran,[51] and Russia[52] have

---

respected." <https://fra.europa.eu/en/eu-charter/article/13-freedom-arts-and-sciences#:~:text=The%20arts%20and%20scientific%20research%20shall%20be%20free%20of%20constraint.>

44 See Leerssen (n 40).

45 For instance, Chander alludes to the possibility of misuse of data access provisions under the DSA in imperfect democracies and authoritarian regimes. Anupam Chander, 'When the Digital Services Act Goes Global' (18 October 2023) <https://papers.ssrn.com/abstract=4606282> accessed 25 October 2023.

46 See for instance, Niha Masih, Shams Irfan, and Joanna Slater, 'India's Internet Shutdown in Kashmir Is the Longest Ever in a Democracy' *Washington Post* (16 December 2019) <https://www.washingtonpost.com/world/asia_pacific/indias-internet-shutdown-in-kashmir-is-now-the-longest-ever-in-a-democracy/2019/12/15/bb0693ea-1dfc-11ea-977a-15a6710ed6da_story.html>; Advait Palepu, 'MEITY Instructs Twitter to Block over 1,000 Accounts Related to Farmers Protest: Report' *MediaNama* (8 February 2021) <https://www.medianama.com/2021/02/223-meity-instructs-twitter-to-block-over-1000-accounts-related-to-farmers-protest-report/> accessed 18 January 2022.

47 See for instance, Human Rights Watch, 'Bangladesh: Internet Ban Risks Rohingya Lives' (26 March 2020) <https://www.hrw.org/news/2020/03/26/bangladesh-internet-ban-risks-rohingya-lives>; 'Bangladesh: Freedom on the Net 2023 Country Report' (*Freedom House*) <https://freedomhouse.org/country/bangladesh/freedom-net/2023> accessed 26 November 2023.

48 See for instance, Frances Mao, 'Pakistan Shut down the Internet - but That Didn't Stop the Protests' (12 May 2023) <https://www.bbc.com/news/world-asia-65541769>; 'Pakistan Says It Blocked Social Media Platform X over "National Security"' *Al Jazeera* (17 April 2024) <https://www.aljazeera.com/news/2024/4/17/pakistan-says-it-blocked-social-media-platform-x-over-national-security>.

suspended the internet during political protests or blocked access to political speech and dissenting content on grounds of national security and public order, disinformation or obscenity.[53] Data access could potentially become another mechanism empowering the executive to indirectly exercise control over online speech.[54] Overbroad research purposes like "public security" or "civic discourse" could be used to justify intelligence gathering by state-aligned researchers or even law enforcement (see Section 5.7) in the absence of adequate safeguards.

Global South countries must, therefore, carefully deliberate whether the scope of research needs to be open-ended or whether it must be linked to risk assessment frameworks while designing their data access legislation. It might be preferable to keep the research scope open-ended, and even in circumstances when it is defined, the scope must reflect their local socio-political contexts while ensuring adequate safeguards against state abuse.

---

49 See for instance, Feranmi Adeoye, 'Issues in Internet Regulation in Nigeria: The Need to Promulgate a Befitting Legislation' (25 March 2020) <https://papers.ssrn.com/abstract=3773010> accessed 5 December 2023; Vincent A. Obia, 'Twitter versus Government of Nigeria: Power, Securitisation and the Politics of a Social Media Ban' (*Media@LSE*, 17 June 2021) <https://blogs.lse.ac.uk/medialse/2021/06/17/twitter-versus-government-of-nigeria-power-securitisation-and-the-politics-of-a-social-media-ban/> accessed 19 September 2022.

50 See for instance, Adam Samson, 'Turkey Tightens Internet Censorship Ahead of Elections' (14 January 2024) <https://www.ft.com/content/c042a067-3cb2-48fe-90f8-e61ee6824b5e> accessed 13 May 2024; Vasilis Ververis, Sophia Marguel and Benjamin Fabian, 'Cross-Country Comparison of Internet Censorship: A Literature Review' (2020) 12 Policy & Internet 450.

51 Ververis, Marguel and Fabian (n 50).

52 See for instance, Alena Epifanova, 'Throttling of YouTube Shows That Russia Is Getting Better at Online Censorship' (*Carnegie Endowment for International Peace*) <https://carnegieendowment.org/russia-eurasia/politika/2025/02/russia-youtube-block-attempt?lang=en>; Paul Mozur, Adam Satariano and Aaron Krolik, 'Russia's Online Censorship Has Soared 30-Fold During Ukraine War' *The New York Times* (26 July 2023) <https://www.nytimes.com/2023/07/26/technology/russia-censorship-ukraine-war.html>.

53 See for instance, Tavishi and others, 'Social Media Regulation and the Rule of Law: Key Trends in Sri Lanka, India and Bangladesh' (CLJ Malaysia SdnBhd 2024) <https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/social-media-regulation-and-the-rule-of-law-ebook-2nd-edn-681.pdf>.

54 See Chander (n 45); Isabelle Canaan, 'NetzDG and the German Precedent for Authoritarian Creep and Authoritarian Learning' (10 August 2021) <https://papers.ssrn.com/abstract=3908440> accessed 14 December 2023.

As regards the territorial scope of research under the DSA, although research is limited to systemic risks in the EU, several scholars argue that this should not limit the geographical scope of data requests.[55] Some suggest that non-EU countries could serve as important control groups for experimental research or even case studies for comparative research.[56] Others argue that understanding systemic risks of, for instance, hate speech, extremist content or disinformation, as they exist in Global South countries, might be useful for researchers in the EU working on similar issues.[57] Edelson et al.[58] suggest that the global nature of the Internet makes it imperative to develop both "objective criteria" (e.g. the number of European users that have engaged with a certain piece of content originating outside the EU) and "subjective criteria" (e.g. linkages to systemic risks in the EU) to evaluate data access requests on a case to case basis. Here, it might be relevant to point out that social science research might benefit from examining the interconnectedness of global narratives and conspiracy theories that are often cross-border in their nature and impact.[59]

Undoubtedly, access to platform data beyond the EU will strengthen platform transparency globally and also provide some relevant insights for the Global South. However, it is important to consider the implications of data access regimes that

55 Paddy Leerssen, 'Digital Services Act: Summary Report on the Call for Evidence on the Delegated Regulation on Data Access' (EC 2023) <https://digital-strategy.ec.europa.eu/en/library/digital-services-act-summary-report-call-evidence-delegated-regulation-data-access>.

56 In this context Husovec suggests that 'geographical limitation should apply to the main research question and not the research methodology'. See Martin Husovec, 'How to Facilitate Data Access under the Digital Services Act' (2023) <https://ssrn.com/abstract=4452940> accessed 21 June 2023.

57 University of Oxford, 'Written Evidence Submitted by Lujain Ibrahim, Dr Luc Rocher, Dr Ana Valdivia, University of Oxford' (EC 2023) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3423742_nl>.

58 Laura Edelson, Inge Graef, and Filippo Lancieri, 'Access to Data and Algorithms: For an Effective DMA and DSA Implementation' (Centre on Regulation in Europe (CERRE) 2023) <https://www.politico.eu/wp-content/uploads/2023/03/15/230223_Access-to-Data-Algorithms.pdf>.

59 See for instance, 'How France's "Great Replacement" Theory Conquered the Global Far Right' (*France 24*, 8 November 2021) <https://www.france24.com/en/europe/20211108-how-the-french-great-replacement-theory-conquered-the-far-right> accessed 13 May 2024.

focus exclusively on risks in the EU, while similar data access opportunities are lacking for Global South researchers. This scheme of data access and platform research can reproduce historical power hierarchies that exist between the "European researcher" and the "researched subject located in the Global South".[60] This epistemic inequality and Eurocentrism[61] can result in risk mitigation strategies for platform harms, in the form of new regulations or platform design, that are geared disproportionately towards harms observed in the Global North.

Although stakeholders have recommended that "vetted researchers" should not be restricted to those residing in the EU,[62] it remains to be seen how many researchers from the Global South gain data access through this mechanism. Such researchers face many barriers that hinder their participation in processes centred in the EU, including resource and funding constraints.[63] Academics and institutions are often pressed to prioritise their local and regional research agendas within the limited resources available, especially as comparative studies remain under-funded in most of the region.[64]

Moreover, international collaborative research on platform data, or any research relying on data from other jurisdictions (EU researchers accessing data beyond its borders or non-EU researchers accessing EU data), need to contend with legal prerequisites for data transfers. They must also consider inter-jurisdictional legal conflicts in terms of data protection laws, competition laws, and ethical frameworks

60 Edward W Said, Orientalism: Western Conceptions of the Orient (Penguin UK 2016).

61 George Gheverghese Joseph, Vasu Reddy and Mary Searle-Chatterjee, 'Eurocentrism in the Social Sciences' (1990) 31 Race & Class 1 <http://journals.sagepub.com/doi/10.1177/030639689003100401> accessed 8 May 2024.iv

62 Paddy Leerssen (n 55).

63 Agustina Del Campo, 'Challenges to Majority World Participation in European Union's Data Access for Platform Researchers Consultation' <https://observatoriolegislativocele.com/challenges-to-global-south-participation-in-european-unions-data-access-for-platform-researchers-consultation/>.

64 ibid.

for research.[65] Lenhart, thus, points to the importance of multilateral agreements to facilitate international research collaborations.[66]

Even if researchers in the Global South, through collaboration with EU researchers and institutions, gain access to platform data under the DSA, the scope of research will be centred on risks in the EU. Thus, unless similar data access regimes are instituted in Global South jurisdictions, they will have to be dependent on voluntary mechanisms offered by platforms or independent data collection methods, as seen in Section 2.

5.5. Data Access to Vetted Researchers

In order to gain approval for data access under Article 40(4), researchers must submit an application to the DSC of the establishment[67] or the DSC of the member state in which the affiliate research organisation is located.[68] The DSA lays down the conditions that must be met by applicants to be granted the "vetted researchers" status.[69] Once an application is approved, the DSC of establishment initiates a reasoned request to the platforms to provide data access to the vetted researchers. The DSCs are also empowered to terminate data access when it determines that the vetted researcher no longer meets the baseline conditions.[70]

65 Anna Lenhart, 'A Vision for Regulatory Harmonization to Spur International Research' (*Lawfare*, 3 May 2023) <https://www.lawfaremedia.org/article/a-vision-for-regulatory-harmonization-to-spur-international-research> accessed 24 July 2024.

66 ibid.

67 Digital Services Coordinator of establishment is defined under article 3(n) of the DSA as the Digital Services Coordinator of the Member State where the main establishment of a provider of an intermediary service is located or its legal representative resides or is established

68 According to article 40(9) researchers can also submit their application to the DSC of the Member state of the research organization. This DSC will then conduct an initial assessment and submit the application with the assessment to the DSC of establishment who shall make the final decision on granting "vetted researcher" status.

69 DSA 2022, art 40(8).

70 DSA 2022, art 40(10).

a. Vetting of Researchers

Arriving at reasonable qualifying criteria for researchers is key to successful implementation. On one hand, maintaining robust standards is necessary to ensure data security, confidentiality and privacy; on the other, an excessively high threshold can result in the exclusion of researchers.

To qualify as "vetted researchers" under the DSA, researchers must fulfil the following conditions:[71] (a) be affiliated with a research organisation;[72] (b) be independent of commercial interests; (c) disclose funding; (d) be capable of fulfilling data security and confidentiality requirements; (e) demonstrate that the data access request is necessary and proportionate; (f) their planned research activities should fall within the ambit of the defined research purposes, and (g) they must commit to making their research publicly available free of charge.

The law, thus, restricts data access to researchers affiliated with a research organisation, which is broader than "affiliation to academic institutions" in an earlier draft.[73] The earlier draft was heavily criticised for being too narrow and notably excluding civil society groups.[74] Although the current provision also excludes independent researchers and journalists, this may be a necessary regulatory decision, given the difficulty of regulating and verifying these categories, which makes them more likely to be abused.[75] The question of affiliation is relevant because institutional data management and security practices, and research ethics guidelines might be

71 DSA 2022, art 40(8).

72 Research organisation under Article 40(8)(a) refers to a research organisation within the meaning of Article 2 of Directive (EU) 2019/790 (see recital 97). This includes a "university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research: (a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or (b) pursuant to a public interest mission recognised by a Member State".

73 Council Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and Amending Directive 2000/31/EC [2020] <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A825%3AFIN.>.

74 Frances Haugen, 'Civil Society Must Be Part of the Digital Services Act' *Financial Times* (29 March 2022) <https://www.ft.com/content/99bb6c10-bb09-40c0-bdd9-5b74224a5086> accessed 3 June 2023.

75 See Leerssen (n 40).

essential to safeguard against malicious or inadvertent privacy and security breaches of data. The draft delegated regulation laying down the technical conditions and procedures under which providers of VLOPs and VLOSEs are to share data under article 40 released by the Commission in October 2024 ("Draft Delegated Regulation"),[76] mandates each applicant researcher to provide documentary evidence of a formal relationship with the research organisation of affiliation.[77]

In order to guard against commercial use of data access mechanisms for private interests, the DSA clarifies that research organisations include civil society organisations that conduct scientific research to support their public interest mission.[78] The DSA also requires the applicants to disclose their funding to conduct the research.[79] Several researchers, in their submission to the EC's call for evidence, have pointed out the challenges of getting funding before data access is approved by the DSC.[80] This difficulty persists in the Draft Delegated Regulation, which mandates verification of funding information as a prerequisite for approval of research application by the DSC.[81] This challenge to demonstrate funding at the initial application stage might be magnified multifold for Global South researchers where access to funding is scarce.[82] Edelson et al., in their submission to the EC, have

76 'Draft Commission Delegated Regulation (EU) Supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council by Laying down the Technical Conditions and Procedures under Which Providers of Very Large Online Platforms and of Very Large Online Search Engines Are to Share Data Pursuant to Article 40 of Regulation (EU) 2022/2065' <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act_en>.

77 Draft Delegated Regulation 2024, art 8(2)(b).

78 DSA 2022, recital 97.

79 DSA 2022, art 40(8)(c).

80 See Paddy Leerssen (n 55).

81 Draft Delegated Regulation 2024, art 8(3).

82 For instance, consider the following indicators: As per the data released by the UNESCO Institute For Statistics, the Gross Domestic Expenditure on R&D (GERD) as a percentage of GDP, Northern America spent 3.32% and Europe spent 2.02% of GDP on R&D, while Latin America and the Caribbean spent 0.55%, Southern Asia spent 0.57%, Central Asia spent 0.13%, Northern Africa spent 0.74%, Sub-Saharan Africa spent 0.33% and Eastern Asia spent 2.71% of their GDP in R&D in 2021. See UNESCO Institute For Statistics, 'Science, Technology and Innovation: Gross Domestic Expenditure on R&D (GERD), GERD as a Percentage of GDP, GERD per Capita and GERD per Researcher'<http://data.uis.unesco.org/index.aspx?queryid=74> accessed 16 December 2023.

recommended setting up specialised public funding under "DSA Research Grants".[83] Setting up a similar public fund in Global South countries that bring regulation on researcher access could be a good step to maintain independence of funding from Big Tech. The composition of the authority disbursing such grants would be crucial to maintain the independence of funding.[84]

Importantly, the vetting of researchers under the DSA has been entrusted to DSCs. The DSCs are envisaged as independent bodies with administrative and financial independence guaranteed by law.[85] The DSA also provides for the establishment of independent advisory mechanisms "in support of sharing of data, taking into account the rights and interests of the providers of VLOPs or VLOSEs and the recipients of the service concerned, including the protection of confidential information, in particular trade secrets, and maintaining the security of their service".[86] These independent advisory mechanisms can provide capacity and expertise to the DSCs in vetting a high volume of research applications.[87] Keller, for instance, has argued that the DSCs must not be restricted by the criteria laid out in the DSA to evaluate

83 Laura Edelson, Inge Graef and Filippo Lancieri, 'Response to the European Commission's Call for Evidence for a Delegated Regulation on Data Access Provided for in the Digital Services Act' (Centre on Regulation in Europe (CERRE) 2023) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3422315_en>.

84 Edelson et al. have recommended "specific review committees that incorporate not only other researchers and representatives of the funding agencies, but also representatives of the DSCs and, in a consulting role, representatives of the VLOP/VLOSEs." to manage such funds. See Laura Edelson, Inge Graef, and Filippo Lancieri (n 58); However, other submissions to the EC have warned against the role of DSCs in funding allocation and VLOPs/VLOSEs in research proposal evaluations. See Paddy Leerssen (n 55).

85 DSA 2022, art 50.

86 DSA 2022, art 40(13).

87 Algorithm Watch, 'Call for Evidence: Data Access Rules Must Empower Researchers Where Platforms Won't' (2023) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3423286_en>; Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (2023) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3422727_en>; Stanford Internet Observatory, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (2023) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3423222_en>.

research applications, and should be able to seek other information and consult independent experts on data privacy, data analysis, security, surveillance, etc.[88] It is also suggested that the independent advisory mechanism could provide a "peer review" of requests by leveraging a network of experts in the field of social media research. [89] This mechanism of peer review can also help in preserving the independence of the research agenda.

Thus, the composition of the independent advisory body becomes crucial to maintaining the independence of research. Here, it must be noted that, on one hand, several industry submissions to the EC have recommended including platform representatives in this body, while on the other hand, several research organisations have cautioned against such representation from VLOPs/VLOSEs.[90]

The Draft Delegated Regulation has also affirmed the importance of an independent advisory mechanism and empowered the DSC to consult independent and impartial experts that have no ties with the data provider or the researcher and are free from conflict of interest.[91] The DSC can consult such experts for formulating a reasoned request for data access or for deciding on a data amendment request raised by the platforms.[92]

It is evident that the expertise and independence of institutions (DSCs, independent advisory bodies, public fund managing committees, etc.) tasked with vetting researchers is fundamental to the successful implementation of a data access regime for researchers. This can become a significant challenge in several Global South countries which lack the requisite infrastructure and technical, administrative and

88 Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87).

89 See European Digital Media Observatory and (EDMO), 'Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access' (2022) <https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf>; Stanford Internet Observatory (n 87); Paddy Leerssen (n 55).

90 Paddy Leerssen (n 55).

91 Draft Delegated Regulation 2024, art 14.

92 Draft Delegated Regulation 2024, art 14(1).

financial resources to evaluate research proposals. Moreover, in many jurisdictions, specialised regulators for digital platforms either do not exist or lack independence from the executive[93] and the states often employ regulatory and technical tools to censor online speech.[94] In this context, if vetting of research applications is undertaken by an executive authority, research agendas that are critical of the state or the dominant political ideology might never be approved. Instead, only those researchers who closely follow the priorities of the state or its law enforcement agencies might qualify for data access, which could lead to significant risks of state surveillance and censorship. The absence of independent and representative institutions to vet research applications might also result in a lack of diversity in the research agenda.

The magnitude of the risks associated with potential state control over the selection of researchers becomes evident from the global disparity in academic freedom. Countries located in Asia and the Pacific, the Middle East and North Africa, Sub-Saharan Africa, Eastern Europe and Central Asia, have some of the lowest levels of academic freedom, and many among them have also experienced significant decline

93 For instance, in Turkey, the board members of the Information and Communication Technologies Authority (BTK) are appointed by the government and the regulator lacks independence. See 'Turkey: Freedom on the Net 2022 Country Report' (*Freedom House*) <https://freedomhouse.org/country/turkey/freedom-net/2022> accessed 13 May 2024; In South Asia, India lacks an independent digital regulator for social media platforms, Sri Lanka has passed law to establish the Online Safety Commission which has been criticised for its wide powers and lack of independence. Similarly, Bangladesh's Cyber Security Act had established the National Cyber Security Council headed by the Prime Minister. Tavishi and others (n 53); In a proposed legislation in Nigeria, law enforcement wields excessive discretionary power. See Verengai Mabika, Emmanuel C Ogu and DearGovernments Organization, 'Nigeria's Protection from Internet Falsehood and Manipulation Bill 2019'.

94 See Ververis, Marguel and Fabian (n 50); Zahra Takhshid, 'Regulating Social Media in the Global South' (2021) 24 Vanderbilt Journal of Entertainment & Technology Law 1 <https://scholarship.law.vanderbilt.edu/jetlaw/vol24/iss1/1>; Manish Singh, 'India Blocks YouTube Videos and Twitter Posts on BBC Modi Documentary' (*TechCrunch*, 21 January 2023) <https://techcrunch.com/2023/01/21/india-blocks-youtube-videos-and-twitter-posts-on-bbc-modi-documentary/> accessed 9 October 2023; Asantha Sirimanne and Anusha Ondaatjie, 'Sri Lanka Throttles Social Media, Protests as Unrest Builds' *Bloomberg.com* (3 April 2022) <https://www.bloomberg.com/news/articles/2022-04-03/sri-lanka-blocks-social-media-imposes-curfew-to-curb-protests>; Emmanuel Akinwotu, 'Nigeria Lifts Twitter Ban Seven Months after Site Deleted President's Post' *The Guardian* (13 January 2022) <https://www.theguardian.com/world/2022/jan/13/nigeria-lifts-twitter-ban-seven-months-after-site-deleted-presidents-post> accessed 19 September 2022; Asantha Sirimanne and Anusha Ondaatjie.

over the last decade.[95] Academic unfreedom translates to appointments based on political affiliation, restriction on research and teaching and freedom of academic and cultural expression, among other things.[96] Global South countries with low levels of academic freedom will be at a higher risk of powerful executives extending their control over upcoming avenues of academic research like the data access mechanisms discussed in this chapter.

b. Trade-offs with data privacy and security

In order to qualify for data access, applicants must demonstrate that the access to data and the timeframes are "necessary for, and proportionate to, the purposes of their research".[97] It is important to note that the vetting of research applications must contend with trade-offs between user privacy and data access. This is more likely to be determined on a case-by-case approach, at least until standards evolve.[98] As noted by Keller,[99] different types of data present different levels of privacy risks, and these questions must be carefully considered while deciding on data access modalities. Accessing actual content can often disclose some personally identifying information, but this data may be essential for several researchers, including those trying to gauge the efficacy and fairness of platforms' content moderation

95 Katrin Kinzelbach, Staffan I. Lindberg, and Lars Lott, 'Academic Freedom Index 2024 Update' (FAU Erlangen-Nürnberg and V-Dem Institute 2024) <10.25593/open-fau-405> accessed 13 August 2024; Michael Coppedge et al., 'V-Dem Dataset V13' <https://www.v-dem.net/data/dataset-archive/> accessed 30 May 2023.

96 Nandini Sundar and Gowhar Fazili, 'Academic Freedom In India' (*The India Forum*, 27 August 2020) <https://www.theindiaforum.in/article/academic-freedom-india> accessed 30 May 2023.

97 DSA 2022, art 40(8)(c).

98 Article 8(7) of the Draft Delegated Regulation mandates the research applications contain proposed safeguards to mitigate confidentiality, security and privacy risks for the data access and processing. Article 9 empowers the DSC to determine the modality of data access, including the interface for accessing data (like online databases, APIs, secure processing environment, etc.) and the accompanying legal, organisational and technical conditions for access taking into consideration several factors including the sensitivity of the data and interests of the online platforms.

99 Daphne Keller, 'User Privacy vs. Platform Transparency: The Conflicts Are Real and We Need to Talk About Them' (*The Center for Internet and Society at Stanford Law School*, 6 April 2022) <https://cyberlaw.stanford.edu/blog/2022/04/user-privacy-vs-platform-transparency-conflicts-are-real-and-we-need-talk-about-them-0>.

mechanisms. Even aggregate data or anonymised longitudinal data sets can lead to the reidentification of individuals. Although methods to share privacy-protecting datasets exist, it might be useful for experts to assess which method will be more useful for a particular case. For instance, differential privacy datasets, which add "noise" to real data, might be useful for researchers who have sufficient statistical training and are working on larger user groups.[100] Thus, some recommend a tiered approach to data access, where different mechanisms and safeguards for access apply based on the sensitivity of the data requested.[101]

Another additional aspect of privacy that might assume significance in many Global South countries is with respect to the private messaging services offered by social media platforms. Often, hate speech and disinformation with catastrophic real-life consequences spread on private messaging applications in the Global South. India has witnessed mob lynchings based on rumours spread on WhatsApp.[102] Meanwhile, the proliferation of electoral misinformation and conspiracy theories on platforms like WhatsApp and Telegram culminated in the January 8 Brasília attacks in Brazil.[103] It thus becomes imperative to evaluate if researchers should gain access to metadata of the private messaging services offered by social media platforms. An effective policy must contend with questions of surrounding users' expectations of privacy on private messaging services.[104] These debates also become extremely important given that end-to-end encryption in personal messaging platforms is being

100 ibid.

101 European Digital Media Observatory and (EDMO) (n 89); Mathias Vermeulen (n 3); Paddy Leerssen (n 55).

102 'How WhatsApp Helped Turn an Indian Village into a Lynch Mob' BBC News (18 July 2018) <https://www.bbc.com/news/world-asia-india-44856910>.

103 See Sheera Frenkel, 'The Pro-Bolsonaro Riot and Jan. 6 Attack Followed a Similar Digital Playbook, Experts Say.' *The New York Times* (10 January 2023) <https://www.nytimes.com/2023/01/09/technology/brazil-riots-jan-6-misinformation-social-media.html> accessed 12 March 2025; Joao VS Ozawa and others, 'Brazilian Capitol Attack: The Interaction between Bolsonaro's Supporters' Content, WhatsApp, Twitter, and News Media' [2024] Harvard Kennedy School Misinformation Review <https://misinforeview.hks.harvard.edu/article/brazilian-capitol-attack-the-interaction-between-bolsonaros-supporters-content-whatsapp-twitter-and-news-media/>.ipi

104 Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87).

undermined in some of these countries through mechanisms such as "traceability of the originator of the message".[105]

Even after vetting researchers, it is difficult to enforce privacy rules on large datasets. Here, scholars[106] have highlighted the importance of making data access mechanisms compliant with the General Data Protection Regulation (GDPR) through a code of conduct[107] which can clearly lay down the obligations for DSC(s), platforms and researchers. The European Data Protection Supervisor (EDPS) has also emphasized the importance of codes of conduct to facilitate scientific research.[108] The European Digital Media Observatory (EDMO) has released a draft code that lays down how data-sharing organisations and researchers can delimit their legal roles, responsibilities and liabilities, as well as the relevant exemptions and derogations under GDPR for processing personal data for research and the corresponding safeguards pertaining to transparency and data subject rights.[109]

At this point, it is essential to note that many countries in the Global South have no or weak personal data protection legislation.[110] While on one hand, this can obviously lead to researchers inadvertently exposing users to privacy and security harms, ambiguous data protection legislation which doesn't have express exemptions and

105See John Xavier, 'WhatsApp vs Government | Why Exiting India Threat Bestirs "Traceability" Debate' *The Hindu* (27 April 2024) <https://www.thehindu.com/sci-tech/technology/whatsapp-vs-government-why-exiting-india-threat-bestirs-traceability-debate/article68113037.ece> accessed 13 May 2024; Collaboration on International ICT Policy for East and Southern Africa (CIPESA), 'How African Governments Undermine the Use of Encryption' (2021) <https://cipesa.org/wp-content/files/briefs/How_Africa_Government_Undermine_the_Use_of_Encryption_2021.pdf> accessed 27 August 2024.

106 Mathias Vermeulen (n 3).

107 This code of conduct would be adopted under Article 40 of the GDPR.

108 European Data Protection Supervisor, 'A Preliminary Opinion on Data Protection and Scientific Research' (2020) <https://www.edps.europa.eu/sites/default/files/publication/20-01-06_opinion_research_en.pdf>.

109 European Digital Media Observatory and (EDMO) (n 89).

110 Graham Greenleaf, 'Global Data Privacy Laws 2023: 162 National Laws and 20 Bills' (10 February 2023) <https://papers.ssrn.com/abstract=4426146> accessed 13 May 2024.

safeguards carved out for research can also be used by platforms to deny researchers access to any data.[111]

Apart from the privacy harms, researcher access to platform data comes with significant data security risks. Researchers create new confidentiality risks through inadvertent errors and leaks.[112] There is also the risk of researchers who once gained access to platform data for public interest independent research moving to work in the corporate or government sector, where the insights gained from such access can lead to competitors or regulators gaining knowledge of the internal processes of companies.[113] Several submissions to the EC's call for evidence,[114] suggested data management plans, non-disclosure agreements (NDAs) and activity logging to combat these potential harms. [115]

To summarise, robust data protection frameworks, codified data security practices and guidelines, and institutional oversight to ensure compliance from platforms and researchers, are essential prerequisites for effective data access. This can be an uphill task for many Global South countries where data protection regulation is still at a nascent stage and there are limited resources and capacity for evolving codes of conduct and standards.

c. Operationalising data access

The DSA does not prescribe detailed mechanisms for operationalising data access for researchers. It simply lays down that platforms shall provide access to data

111 Mathias Vermeulen (n 3); Jef Ausloos and Michael Veale, 'Researching with Data Rights' (2020) 2020–30 Amsterdam Law School Research Paper <https://doi.org/10.26116/techreg.2020.010>.

112 Daphne Keller, 'Some Practical Postulates About Platform Data' (18 May 2023) <https://cyberlaw.stanford.edu/blog/2023/05/some-practical-postulates-about-platform-data> accessed 19 May 2023.

113 ibid.

114 Paddy Leerssen (n 55).

115 The Draft Delegated regulation has affirmed the use of such legal, organisational and technical safeguards in vetting of research applications.

See recital 13 of the Draft Delegated Regulation 2024.

requested under Art 40(4) "through appropriate interfaces specified in the request, including online databases and application programming interfaces".[116] As seen in the previous section, the Draft Delegated Regulation empowers the DSCs to determine the appropriate access modality, including an option to create a secure processing environment provided by or on behalf of the online platform.[117] This leaves flexibility for systems to evolve and accommodate rapidly changing technology. However, as per the Draft Delegated Regulation, it appears that the DSC's reasoned request to the platforms, including the access modalities and period for data access, cannot be amended or extended by the researchers at a later stage. This might be challenging for researchers, given the unchartered territory of platform data.

To begin with, researchers will face the challenge of "unknown unknowns", where they have to develop their research questions without sufficient knowledge of the kind of data that can be made available by platforms.[118] Further, as Keller[119] points out, it might even be difficult for researchers to ask for the right data or the correct metrics in the first instance, given how research questions tend to evolve throughout the life cycle of a research project/endeavour. Similarly, platforms are also likely to not have the datasets readily available for all possible research questions and most likely will have to invest resources in pulling the right data.[120] Several submissions by researchers to the EC's call for evidence emphasised the importance of exploratory research in this context.[121] However, industry responses warned against "fishing

116 DSA 2022, art 40(7).

117 Draft Delegated Regulation 2024, art 9.

118 Elizabeth Hansen Shapiro and others, 'New Approaches to Platform Data Research' (Netgain Partnership 2021) <https://www.netgainpartnership.org/resources/2021/2/25/new-approaches-to-platform-data-research>in Stanford Internet Observatory (n 87).

119 Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87); Daphne Keller, 'Some Practical Postulates About Platform Data' (n 112).

120 ibid.

121 Stiftung Neue Verantwortung (SNV), 'Response to the European Commission's Call for Evidence on a Planned Delegated Regulation on Data Access Provided for in the Digital Services Act (DSA)' (2023) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3422376_en>; University of Michigan Center for Social Media Responsibility, 'Response to the European Commission's Call for Evidence for a Delegated Regulation on Data Access Provided for in the Digital Services Act' (EC 2023) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-

expeditions", arguing that data access requests should strictly adhere to what is "necessary for and proportionate to the purpose of the research".[122]

However, it is also likely that platforms might pull imperfect data at first and provide additional data or fixes later.[123] Thus, data access in practical terms can be best understood as an iterative process needing close coordination between platforms, researchers and the DSC to ensure that relevant and complete data is received in accessible formats.[124] This need for close coordination between platforms and researchers becomes even more crucial for designing and conducting experimental studies,[125] or qualitative research relying on access to platforms' internal guidelines and employee interviews, or even accessing information related to internal studies conducted by the platforms themselves.[126]

Several researchers have also highlighted the importance of transparency in platforms' internal documentation, including the codification and classification of data, as platforms and researchers could differ in their understanding of how a particular data attribute is defined.[127] Acknowledging this concern, the Draft Delegated Regulation mandates platforms to provide relevant documentation to the data requested except when such a disclosure could lead to significant vulnerabilities for the platform.[128]

Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3423924_nl> in Paddy Leerssen (n 55).

122 'CCIA, DOT Europe and Booking.com all wish to exclude fishing expeditions from the scope of Article 40'. See Paddy Leerssen (n 55).

123 Daphne Keller, 'Some Practical Postulates About Platform Data' (n 112).

124 Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87).

125 Stiftung Neue Verantwortung (SNV) (n 121); Daria Dergacheva and others, 'Improving Data Access for Researchers in the Digital Services Act' [2023] SSRN Electronic Journal <https://www.ssrn.com/abstract=4465846> accessed 12 October 2023; Husovec (n 56).

126 See Paddy Leerssen (n 55).

127 ibid.

128 Draft Delegated Regulation 2024, art 15(2).

These challenges of operationalising data access are magnified multifold for Global South researchers as they often don't have access to the past experience of Global North researchers. Elite universities in the USA and EU have accumulated some institutional knowledge over time with data-sharing agreements or access to platform APIs and tools,[129] which are often not available to researchers located in the Global South. Moreover, the uneven allocation of platform resources across jurisdictions and the neglect of Global South[130] means that countries which aren't priority markets with large user bases are unlikely to have dedicated personnel to analyse data requests and retrieve relevant data. Platforms may lack staff trained in local laws, socio-political contexts and linguistic diversity, and they may have little incentive to address this gap. This will result in long-term and iterative collaboration with platforms for data access requests even more difficult than in the EU.

Given the imminent challenges in facilitating coordination between platforms and researchers during the early stages of the DSA, some scholars have proposed the creation of standardised, readily accessible topical datasets and APIs for commonly requested data as a starting point for facilitating data access.[131] These datasets could be used by researchers initially, and more custom requests for data could evolve as institutional vetting and access processes mature. This iterative approach could be useful for the Global South, too. It can potentially mitigate against procedural delays and even long litigation battles with platforms paving the way for data access to Global South researchers who can later use their experience to demand more customised data sets.

---

129 'Status Report: Mechanisms for Researcher Access to Online Platform Data' (n 11).

130 Ben Gilbert, 'Facebook Ranks Countries into Tiers of Importance for Content Moderation, with Some Nations Getting Little to No Direct Oversight, Report Says' *Business Insider* (5 October 2021) <https://www.businessinsider.in/tech/news/facebook-ranks-countries-into-tiers-of-importance-for-content-moderation-with-some-nations-getting-little-to-no-direct-oversight-report-says/articleshow/87263447.cms> accessed 17 May 2023; Cat Zakrzewski and others, 'How Facebook Neglected the Rest of the World, Fueling Hate Speech and Violence in India' *Washington Post* (24 October 2021) <https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/>.

131 Edelson, Graef and Lancieri (n 83); Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87).

This tiered data access can also incorporate the risk-based approach evaluating considerations of both researcher accessibility and efficiency, as well as, data sensitivity and security in determining the mode of data access.[132]

Another important suggestion to facilitate better coordination between platforms and researchers is setting up an independent body for aiding the DSCs in evaluating research proposals and operationalising access.[133] The independent body could also undertake auditing of data access processes and datasets so that the quality of data can be ensured and substantial lapses or errors can be detected before they alter research outcomes. The body could also document the information available for research requests and details of accessing it, and also establish formatting standards and minimum requirements for data.[134]

To this effect, an EDMO working group comprising members from academia, civil society, and VLOPs/VLOSEs was constituted to discuss the organisational structure and core functions of such an independent intermediary body (IIB).[135] In a report, the working group outlined core tasks and principles for the IIB, including laying down standards and processes for vetting researchers and evaluating proposals, and providing accreditation to other organisations to conduct such assessments.[136] As per the report, the IIB should also develop common standards for data codebooks and conduct audits to check the quality of data shared by platforms, as well as, act as a forum to mediate disputes between platforms and researchers.[137]

---

132 Stanford Internet Observatory (n 87).

133 Daphne Keller, 'Some Practical Postulates About Platform Data' (n 112); Mathias Vermeulen (n 3); Algorithm Watch (n 87); European Digital Media Observatory and (EDMO) (n 89).

134 Stanford Internet Observatory (n 87).

135 EDMO, 'Launch of the EDMO Working Group for the Creation of an Independent Intermediary Body to Support Research on Digital Platforms' (15 May 2023) <https://edmo.eu/edmo-news/launch-of-the-edmo-working-group-for-the-creation-of-an-independent-intermediary-body-to-support-research-on-digital-platforms/> accessed 5 August 2024.

136 EDMO and Institute for Data, Democracy & Politics, The George Washington University, 'Core Tasks and Principles for an Independent Intermediary Body That Will Facilitate Researchers' Access to Platform Data' (2023) <https://edmo.eu/wp-content/uploads/2023/09/Creating-an-Independent-Intermediary-Body-to-Facilitate-Platform-Research69.pdf> accessed 5 August 2024.

137 ibid.

To develop standards and guidelines for vetting researchers, operationalising data access and mediating disputes between researchers and platforms, Riley and Ness have proposed a "modular" approach to data access through a multistakeholder international body.[138]

These multistakeholder intermediary bodies could also benefit researchers in jurisdictions where legal provisions and regulatory capacity to mandate and enforce data access have not yet been developed.[139] The development of common principles and guidelines for research ethics and data practices, as well as, the existence of independent intermediary bodies facilitating such access, could be beneficial to researchers in the Global South too, including those who rely on platforms' voluntary data access mechanisms. However, these bodies must evolve beyond mere representation of stakeholders from the Global South to creating deliberative democratic spaces that foster meaningful participation from a diversity of communities. Moreover, the high-level guidelines and standards evolved in such fora must also leave sufficient room for interpretation and flexibility based on local contexts and socio-economic realities.

d. Platforms' amendment to data access request

Platforms can request an amendment to the data access request received from the DSC within 15 days of receiving it. This can be done in cases where platforms believe they will not be able to provide access to requested data because (a) they do not have access to the data requested or (b) access can result in security vulnerabilities or can breach confidential information, particularly trade secrets.[140] These requests for amendments must suggest alternate means to provide access to requested data or

138 Riley and Ness outline how modularity, 'works by identifying tasks common to laws in multiple countries and creating global, multi-stakeholder processes and institutions that can operationalize those tasks.' See Chris Riley and Susan Ness, 'A Module Playbook for Platform-to-Researcher Data Access' (*Tech Policy Press*, 20 November 2022) <https://techpolicy.press/a-module-playbook-for-platform-to-researcher-data-access> accessed 24 July 2024.

139 ibid.

140 DSA 2022, art 40(5).

suggest other data that would fulfil the purpose of the request.[141] The DSC of the establishment is entrusted with the decision on the platform's amendment request.

These provisions have been criticised for being broad enough to provide platforms sufficient room to deny requests that are antithetical to their interests.[142] The clause on the protection of "confidential information" is ambiguous and broad enough to stall any meaningful transparency. This is in sharp contrast with the access rights granted to auditors, who are not denied access to platform information but are bound to guarantee the confidentiality of trade secrets. Vermeulen argues that similar requirements could have been imposed on researchers instead of providing a blanket exemption to platforms.[143] Platforms might also use this clause to engage in long legal battles, which might delay and ultimately frustrate important research proposals. This challenge gets magnified in Global South countries where researchers might not have the resources to engage in litigation with platforms that have deep pockets.

Creating independent fora for mediation could potentially aid in resolving some of these disputes. For instance, the EDMO working group has recommended that the independent intermediary body function as a forum to mediate disputes between platforms and researchers.[144]

The Draft Delegated Regulation has laid down guidelines for the DSCs to evaluate the amendment request raised by the platforms[145] and provided an independent dispute settlement through mediation.[146] It, however, only empowers the platforms to

141 DSA 2022, art 40(6).

142 Leerssen (n 40).

143 Mathias Vermeulen (n 3).

144 EDMO and Institute for Data, Democracy & Politics, The George Washington University (n 136).

145 See Draft Delegated Regulation 2024, art 12. For instance, in order to decide on an amendment request pursuant to Article 40(5)(b) of the DSA, the DSC must take into account: (a) if the alleged vulnerability raised by the platform and its significance is duly substantiated, (b) the likelihood and severity of harm that can result from such a vulnerability, and (c)the extent to which the access modalities mitigate against the vulnerability.

146 See Draft Delegated Regulation 2024, art 15.

initiate the mediation if they disagree with the decision of the DSC on the amendment request. The DSC can include the researcher as a party to the mediation where it deems appropriate, but researchers cannot initiate the mediation process.

5.6. Access to Public Data

Article 40(12) mandates VLOPs/VLOSEs to provide researchers access to public data, including real-time data, wherever technically feasible, without undue delay. This access is guaranteed to a larger pool of researchers than mandated under Article 40(4) (access to vetted researchers). These researchers include those affiliated with not-for-profit bodies, organisations and associations that are independent of commercial interests,[147] disclose sources of funding,[148] are capable of fulfilling data security and confidentiality [149] and their data use request is necessary and proportionate.[150] Such research must be undertaken for the purpose of detection, identification and understanding of systemic risks identified under Article 34(1).[151]

The DSA itself does not lay down the details of how such access to public data must be operationalised, and delegated legislation could provide more clarity. Several submissions to the EC's call for evidence envisage Article 40(12) as:[152] (a) obligation on platforms to provide reliable access to public data through API(s) and other tools like CrowdTangle and (b) safeguarding rights of researchers to access public data through independent data collection methods like scraping [153] or sock puppet auditing.[154]

147 DSA 2022, art A(8)(b).

148 DSA 2022, art 40(8)(c).

149 DSA 2022, art 40(8)(d).

150 DSA 2022, art 40(8)(e).

151 DSA 2022, art 40(12).

152 Paddy Leerssen (n 55).

153 Luscombe et. al. refer to web or data scraping as "automated extraction of information online" … "so long as information appears on a website, whether textual, auditory, or visual, it can in principle be accessed via web scraping." See Alex Luscombe, Kevin Dick and Kevin Walby, 'Algorithmic

In the past, public interest research that relied on access to public data through platforms' voluntary APIs and data scraping has been instrumental in holding platforms accountable and uncovering significant harms despite the precarious terms set by platforms.[155] Independent data collection methods like scraping and data donation became all the more prominent after many platforms restricted access to their public APIs.[156]

However, platforms have often opposed independent data collection research projects as potential violations of their Terms of Service, privacy and data protection obligations, intellectual property, etc.[157] Meta is known to engage in lawsuits against data scraping [158] and has infamously stalled NYU Ad Observatory [159] and AlgorithmWatch's Instagram monitoring project.[160]Besides legal routes, platforms might impose technical barriers to such research, which are even more difficult to

---

Thinking in the Public Interest: Navigating Technical, Legal, and Ethical Hurdles to Web Scraping in the Social Sciences' (2022) 56 Quality & Quantity 1023 <https://link.springer.com/10.1007/s11135-021-01164-0> accessed 6 May 2024.

154 Sandvig et. al. define, "A sock puppet audit is essentially a classic audit study but instead of hiring actors to represent different positions on a randomized manipulation as testers, the researchers would use computer programs to impersonate users, likely by creating false user accounts or programmatically-constructed traffic." See Christian Sandvig and others, 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms'.

155 Athanasios Andreou and others, 'Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations' (2018); Chengcheng Shao and others, 'The Spread of Fake News by Social Bots' (2017) 96 arXiv preprint arXiv:1707.07592 104; Leerssen, Heldt and Kettemann (n 3).

156 Deen Freelon, 'Computational Research in the Post-API Age' (2018) 35 Political Communication 665 <https://doi.org/10.1080/10584609.2018.1477506> accessed 3 June 2023.

157 Leerssen, Heldt and Kettemann (n 3).

158 Marissa Newman, 'Meta Was Scraping Sites for Years While Fighting the Practice' Bloomberg.com (2 February 2023) <https://www.bloomberg.com/news/articles/2023-02-02/meta-was-scraping-sites-for-years-while-fighting-the-practice>.

159 Horwitz (n 24).

160 Nicolas Kayser-Bril, 'AlgorithmWatch Forced to Shut down Instagram Monitoring Project after Threats from Facebook' (AlgorithmWatch, 13 August 2021) <https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/>.

dismantle and effectively empower platforms' status as gatekeepers of research, in addition to their role as gatekeepers of online speech.[161]

Article 40(12) of the DSA thus provides an unprecedented opportunity to not only empower researchers via regulation of public APIs offered by platforms, but also through the legal protection of independent data collection methods like data scraping. This provision of the DSA has been characterised by Keller as an "effective backstop of the DSA" for being "open-ended and forward-looking".[162]

This is because data scraping is considered a rare form of transparency since it does not need platforms to act as gatekeepers in providing any information. Hence, it doesn't have the scope for errors or deliberate manipulation of data. It essentially means "researchers are able to view content as it exists".[163]

However, it must be noted that independent methods of data collection and access to APIs or tools also produce very real risks in terms of privacy, and these methods can be abused for commercial or political interests.[164] It thus becomes imperative that questions regarding what constitutes public data, mechanisms for granting access to platform APIs/ tools and codes of conduct for independent data collection be deliberated and clearly outlined.

For instance, a group of civil society organisations led by the Mozilla Foundation have provided recommendations for the operationalisation of such public data access.[165] They suggest that data should be complete, comprehensive, and verifiable and must include historical data. Platforms must not hinder independent public interest research, and access must be facilitated through fair and reasonable terms to a diversity of researchers, including journalists and those residing outside the EU.

---

161 Leerssen, Heldt and Kettemann (n 3).

162 Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87).

163 ibid.

164 Leerssen, Heldt and Kettemann (n 3).

165 'The Digital Services Act Must Ensure Public Data for Public Interest Research' (*Mozilla Foundation*, 31 May 2023) <https://foundation.mozilla.org/en/blog/the-digital-services-act-must-ensure-public-data-for-public-interest-research/> accessed 3 June 2023.

Further, several submissions to the EC's call for evidence suggest that the delegated regulation on researcher access must explicitly state that researchers complying with Article 40(12) and scraping platform data for privacy-compliant public-interest research must have legal immunity.[166] Researchers have sought clarity on copyright exemptions,[167] and some have also suggested a positive obligation on platforms to remove technical barriers to public interest research through scraping and other data collection methods.[168]

However, Keller rightly cautions against mobilising Article 40(12) of the DSA to gatekeep data scraping.[169] The DSA only imposes a positive obligation on platforms to facilitate access to public data for a set of researchers. It by no means prohibits data scraping for researchers outside the scope of Article 40(12). Thus, the independent regulatory bodies envisaged under the DSA must not gatekeep but only support researchers and identify them for protection. The larger practice of data scraping should be regulated by data protection legislation since public data contains significant levels of personal information. In the context of the EU, the GDPR and its codes of conduct should provide safeguards and best practices for data scraping in research.[170]

Mandating researcher access to public data through API(s), and providing legal immunity to data scraping for public-interest research are low-hanging fruits that can immensely benefit Global South countries. These do not require complex researcher vetting mechanisms and independent bodies to facilitate long-term researcher-platform collaboration. Also, platforms have traditionally provided public APIs as voluntary mechanisms, and providing mandated and reliable APIs would not

166 See Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87); Paddy Leerssen (n 55); Dergacheva and others (n 125).

167Dergacheva and others (n 125).

168ibid; Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87); Husovec (n 56).

169 Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87).

170ibid; Mathias Vermeulen (n 3); Husovec (n 56).

require reinventing the wheel.[171] Further, Global South countries that lack advanced transparency legislation can start by providing legal immunity to data scraping for public-interest research with sufficient safeguards for user privacy.[172] This can be a good starting point to catalyse research, and once researchers and institutions develop significant expertise, data access mechanisms similar to those under Article 40(4) can be introduced.

However, as mentioned previously, the lack of adequate personal data protection legislation in several countries might expose citizens to potential privacy harms and abuse. Another challenge is the potential misuse of public APIs meant for researchers being used as tools for law enforcement to monitor and surveil citizens (see Section 5.7).[173]

5.7. Other Risks and Challenges of Researcher Access to Platform Data

a. Law Enforcement gaining access to researcher data

Another significant risk is the threat that researcher access to data will be misused for state surveillance. Law enforcement agencies (LEAs) across the globe are

171 This is not to suggest that regulation of APIs does not present its unique challenges that need to be understood and addressed as new regulations and practices evolve. See Leerssen, Heldt and Kettemann (n 3); MZ van Drunen and A Noroozian, 'How to Design Data Access for Researchers: A Legal and Software Development Perspective' (2024) 52 Computer Law & Security Review 105946 <https://www.sciencedirect.com/science/article/pii/S026736492400013X> accessed 8 February 2024.

172 See Leerssen et. al. for instance, "Perhaps the most feasible approach, at least in the short term, might be to develop certification schemes or safe harbors to protect independent scraping efforts from restrictive platform policies". Leerssen, Heldt and Kettemann (n 3); Also see Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87); Husovec (n 56).

173 CDT Europe, 'CDT Europe Contribution to European Commission Public Consultation: Draft Implementing Regulation Laying down Templates Concerning the Transparency Reporting Obligations under the Digital Services Act' (2024) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-/A>.

interested in acquiring social media data.[174] Vogus outlines how LEAs can get access to social media data held by researchers, either through voluntary disclosure in response to informal requests or compelled disclosure under a legal order.[175] Alternatively, researchers might find objectionable content and voluntarily report it to LEAs.[176] The rules governing such interaction and notification need to be clearly laid out in data protection and researcher access regimes. LEAs can also directly gain access to social media data meant for researchers by (a) using mechanisms of data access like public data API(s) or tools meant for researchers or (b) through legal orders to platforms to provide datasets used by researchers that might previously not exist in a useful form or were previously not known to LEAs.[177] Several experts have also recognised that there is a risk of LEAs gaining access to vetted researcher status directly or through a consortium of researchers.[178] There is also the risk of LEAs influencing research agendas based on their specialised research organisations getting vetted status.[179]

174 Caitlin Vogus, 'Defending Data: Privacy Protection, Independent Researchers, and Access to Social Media Data in the US and EU' (Center for Democracy & Technology (CDT) 2023) <https://cdt.org/wp-content/uploads/2023/01/2023-01-27-CDT-Defending-Data-Privacy-Protection-Independent-Researchers-and-Access-to-Social-Media-Data-final.pdf> accessed 5 June 2023; This trend is also visible in India where the central government recently faced a backslash for its plans to develop systems to track sentiments of citizens online. Simultaneously, several state governments have set up social media monitoring cells for law enforcement. See Soumyarendra Barik, 'The Government Wants to Surveil Social Media Users, and Track Their "Sentiments"' *MediaNama* (8 October 2020) <https://www.medianama.com/2020/10/223-india-social-media-surveillance/> accessed 19 July 2022; SELVARAJ A, 'Tamil Nadu: Special Police Unit to Monitor Fake Social Media Posts' *The Times of India* (19 March 2022) <https://timesofindia.indiatimes.com/city/chennai/tamil-nadu-special-police-unit-to-monitor-fake-social-media-posts/articleshow/90315927.cms> accessed 1 October 2022.

175 Caitlin Vogus (n 174).

176 ibid.

177 ibid.

178 In this context, EDMO recommends vetted research entities must not carry out any of the following functions: (i) Law enforcement; (ii) Intelligence services; or (iii) Defence, promotion or upholding of national security. Vogus also recommends not allowing LEA to qualify as vetted researchers. See European Digital Media Observatory and (EDMO) (n 89); Caitlin Vogus (n 174).

179 ibid.

Although there are legitimate public safety interests for law enforcement to access social media data, there is also a scope of abuse where such data is used to target dissenting citizens, especially those belonging to historically marginalised identities. The extent of this threat is highly dependent on the country's legal frameworks under which LEAs can gain access to information, the safeguards in place and their practical implementation. While the EU has significant legal safeguards in place to prevent researcher access from being used as a tool by LEAs, some scholars have noted that there exists ambiguity in law and potential for LEAs to access social media data more easily after disclosure to researchers in the USA.[180]

This threat becomes magnified in several Global South countries where the executive exercises wide discretion in the absence of adequate checks and balances.[181] For instance, in India, LEAs can issue orders without judicial warrants to access data held by intermediaries or, in this case, even researchers to investigate a crime.[182] Also, authorised security agencies can issue confidential orders to intercept, monitor or decrypt any information "generated, transmitted, received or stored in any computer resource" on wide grounds[183] without ex-ante judicial authorisation. Failing to assist state agencies is punishable with imprisonment of up to seven years.[184] Other South Asian countries like Bangladesh[185] and Sri Lanka[186] also have wide monitoring and

180 Caitlin Vogus (n 174).

181 Tavishi and others (n 53).

182 As per Section 94 of the Bharatiya Nagarik Suraksha Sanhita2023, any officer in charge of a police station may issue written orders for production of "*electronic communication, including communication devices, which is likely to contain digital evidence necessary or desirable for the purposes of any investigation, inquiry, trial or other proceeding under this Sanhita.*"

183 Section 69 of the IT Act empowers authorized state officers to issue interception, monitoring and decryption orders "if satisfied that it is necessary or expedient to do in the interest of the sovereignty or integrity of India, defence of India, security of the State, friendly relations with foreign States or public order or for preventing incitement to the commission of any cognizable offence relating to above or for investigation of any offence".

184 The Information Technology Act 2000, s 69(4).

185 Bangladesh Telecommunication Act 2001, s 97(A).

186 Sri Lanka Telecommunications Act 1991, s 54(3).

interception powers with the executive with little safeguards.[187] In Southeast Asia, Malaysia, recently, amended its Communications and Multimedia Act (CMA) to broaden the powers of LEAs to compel disclosure of data for investigative purposes without adequate safeguards.[188] Similarly, in Africa, countries like Egypt provide wide interception powers to national security agencies, and even in jurisdictions where certain legal safeguards exist (like Kenya and South Africa), there have been instances of misuse of the state surveillance infrastructure targeting civil society and journalists.[189]

b. Contributing to existing exploitative systems of surveillance capitalism

While data access for researchers provides a unique framework for platform transparency, Leerssen notes that it does not challenge existing legal structures of trade secrets or terms of service, which have been the bedrock of huge power asymmetry between users and platforms.[190] In fact, the DSA upholds trade secrets as valid grounds for platforms to deny access to data. This raises questions on whether any long-term meaningful accountability that pushes platforms to change basic design and structure would be possible through such mechanisms.

Instead, these mechanisms could have the negative externality of reinforcing and contributing to "surveillance capitalism", as pointed out by Keller.[191] Platforms, for instance, might find ways to monetise the data they collate and organise for researchers through advertisement targeting or other models.[192]

187 Tavishi and others (n 53).

188 'Malaysia: CMA Amendments Are a Step Backwards for Freedom of Expression' (*ARTICLE 19*, 10 December 2024) <https://www.article19.org/resources/malaysia-the-passing-of-the-cma-amendments-is-another-step-backwards-for-freedom-of-expression-joint-statement/>.

189 Tony Roberts and others, 'Surveillance Law in Africa: A Review of Six Countries' (Institute of Development Studies 2021) <https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/16893>.

190 Leerssen (n 40).

191 Daphne Keller, 'Delegated Regulation on Data Access Provided for in the Digital Services Act' (n 87).

192 ibid.

Even though there exist legitimate risks with data access, mandating researcher access to platform data is a significant step forward in ensuring platform accountability and will open platform data for external academic scrutiny for the first time. This provides an immense and unprecedented opportunity to study the risks associated with existing content moderation, recommender systems, and advertising models, as well as possible design and regulatory interventions to tackle online harms.

Insights for the Global South

Data access for research is one of the most promising transparency mechanisms in the DSA. It will unlock platform data for independent public-interest expert scrutiny for the first time. This presents an immense opportunity to critically examine the online information ecosystem and platforms' moderation and curation of user-generated content and advertisements. As a result, data access for research is extremely valuable for countries located outside the EU too, including those in the Global South, where the information asymmetry is even more stark.

However, effectively operationalising complex data access mechanisms, like several other provisions in the DSA, requires strong societal structures, including communities of researchers, empowered civil society actors, and a favourable economic, political and regulatory environment to ensure free, independent and impactful research.[193]

Many countries across the Global South face several challenges to this effect:

- It appears that the power asymmetry between States and Big Tech, as well as Big Tech and Global South researchers, might hinder the ability of most Global South countries to mandate and operationalise such a complex and resource-intensive transparency mechanism in the near future.

- Researcher access to data, as mandated under Article 40(4), comes with a significant regulatory burden. This includes vetting researchers and research applications, as well as determining the modalities for meaningful data access. This requires independence, expertise, infrastructure, and technical, administrative and financial resources, which can prove to be a challenge for many regulators in the Global South at the moment.

- Data access for public interest research requires independent and bipartisan vetting of researchers. This could be a challenge in several Global South

193 See Martin Husovec, 'Will the DSA Work?' [2022] Verfassungsblog<https://verfassungsblog.de/dsa-money-effort/> accessed 14 August 2024.

countries which do not have independent digital regulators and where the executive wields discretionary power to regulate platforms and online speech.

- In many Global South countries, there is a considerable risk that law enforcement agencies could gain access to APIs and tools intended for researchers or obtain the data collected by researchers. This poses serious concerns about privacy violations and the potential for increased surveillance.

- The declining academic freedom in several Global South countries can be a significant challenge in maintaining the independence of the research agenda and ensuring the safety of researchers. This also means that the scope of research must be carefully deliberated to prevent it from being dominated by state interests.

- The absence or inadequacy of data protection legislation in several Global South countries can impact both the privacy of users and the ability of researchers to gain access to platform data. It is essential to have privacy and data protection legislation with derogations for public-interest research and codes of conduct for ethical and privacy-protecting research practices.

- Inadequate funding and infrastructure for data processing, lack of data management and analysis skills, and insufficient institutional support in terms of ethics codes and data security codes might be challenging for several Global South researchers. Thus, allocating public funds for research and capacity building, as well as establishing institutional collaborations with Global North research organisations, could be beneficial.

- Several submissions to the EC have suggested that vetted researchers should not be restricted to those residing in the EU. Similarly, platforms must make APIs and tools under Article 40(12) available to researchers beyond the EU. This can pave the way for Global South researchers to collaborate with institutions and researchers in the EU to study platform data. However, Global South researchers are likely to face several barriers, including resource and funding constraints, as well as inter-jurisdictional legal conflicts limiting data transfers. It is also important that participation from Global South

researchers in international collaborations must go beyond mere representation and be equal and meaningful for all researchers.

As a starting point, researchers in Global South can be provided with mandated access to (i) public data through API(s) and tools and (ii) legal immunity for independent data collection methods like data scraping for public interest research. Although these mechanisms also present challenges pertaining to data privacy and state surveillance, countries can aim to build robust legislation, safeguards, codes of practice and independent bodies. For the long-term, mandated researcher access, similar to that envisioned in Article 40(4), can be pursued. Starting with access to standardised datasets, this can progress to custom data demands as institutions mature and researchers gain more experience and skills.

6. TRANSPARENCY IN CONTENT MODERATION

6.1. Introduction

Platforms play an increasingly important role in determining what speech remains online.[1] However, there has been limited public information on how platforms undertake content moderation.[2] Recently, the proliferation of violent and extremist content, child sexual abuse material (CSAM), non-consensual intimate images (NCII), hate speech, and disinformation has raised concerns about platforms not doing enough to reign in such harmful content.[3] Keller notes that while, on the one hand, platforms have been criticised for not doing enough to remove such harmful speech, there have also been instances of massive public fallouts and political backlash at certain content takedown decisions.[4]

This is because content moderation is inherently political and demands carefully balancing competing interests, values and rights. However, the obfuscation of content moderation systems, especially with the increasing adoption of algorithmic

1 Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Colum. L. Rev. 2011 <https://columbialawreview.org/content/free-speech-is-a-triangle/>.

2 Daphne Keller and Paddy Leerssen, 'Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation' (2020) 220 Social media and democracy: The state of the field and prospects for reform 224.

3 See Kari Paul, 'Facebook Faces Advertiser Revolt over Failure to Address Hate Speech' *The Guardian* (22 June 2020) <https://www.theguardian.com/technology/2020/jun/22/facebook-hate-speech-advertisers-north-face> accessed 16 May 2024; Jon Henley, '85% of People Worry about Online Disinformation, Global Survey Finds' *The Guardian* (7 November 2023) <https://www.theguardian.com/technology/2023/nov/07/85-of-people-worry-about-online-disinformation-global-survey-finds> accessed 17 May 2024; Chad De Guzman and Will Henshall, 'As Tech CEOs Are Grilled Over Child Safety Online, AI Is Complicating the Issue' *TIME* (31 January 2024) <https://time.com/6590470/csam-ai-tech-ceos/> accessed 17 May 2024.

4 Daphne Keller, 'Internet Platforms: Observations on Speech, Danger, and Money' (13 June 2018) <https://papers.ssrn.com/abstract=3262936> accessed 24 September 2022.

moderation,[5] creates an "operating logic of opacity" around moderation decisions by platforms.[6] This, in effect, leads to the illusion of depoliticisation of content moderation.[7] As platforms evolve a labyrinth of opaque processes, technical systems, and exploitative labour practices while being driven by profit motives, they risk replicating existing power structures at the cost of marginalised voices that challenge the status quo.[8]

This lack of transparency and its impact on online speech becomes even more pronounced in Global South countries since platforms allocate little resources to curtail online harm outside their priority markets.[9]

In the previous chapters, we have examined transparency for advertisement models (see Chapter 2), risk assessments (see Chapter 3), audits (see Chapter 4), and researcher access to platform data (see Chapter 5). Transparency in content moderation is additionally operationalised through the following mechanisms under the DSA: (a) through aggregate transparency reporting and disclosure of qualitative information on automated systems and human moderators; (b) through disclosures relating to the Terms and Conditions (T&Cs) based on which platforms govern online speech; (c) through notifications to individual users whose content is actioned by platforms due to violation of the T&Cs or applicable laws.

---

5 Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945 <https://doi.org/10.1177/2053951719897945> accessed 16 March 2022.

6 Sarah T Roberts, 'Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation' [2018] First Monday <https://firstmonday.org/ojs/index.php/fm/article/view/8283>.

7 ibid.

8 ibid.

9 See for instance, the Facebook files revealed that its employees spent only 13% of the total time spent on misinformation outside the US. 'The 5 Most Important Revelations From the "Facebook Papers"' (Time, 25 October 2021) <https://time.com/6110234/facebook-papers-testimony-explained/> accessed 16 December 2023. Also see, Giovanni De Gregorio and Nicole Stremlau, 'Inequalities and Content Moderation' (2023) 14 Global Policy 870 <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.13243> accessed 7 February 2024; Gabriel Nicholas and Aliya Bhatia, 'Toward Better Automated Content Moderation in Low-Resource Languages' (2023) 2 Journal of Online Trust and Safety <https://www.tsjournal.org/index.php/jots/article/view/150> accessed 2 September 2024.

6.2. Mandatory Transparency Reporting

In the 2010s, platforms began issuing transparency reports voluntarily in response to growing demands for accountability regarding state requests for user data and content removal.[10] Snowden's 2013 revelations of surveillance by the US National Security Agency proved to be a significant impetus for platforms to start publishing aggregate information on state requests for user data.[11] While early reports were limited to state requests, more recently, in the aftermath of the 2016 US Presidential Elections, platforms have begun including information on their voluntary content moderation initiatives.[12] Platforms typically provide aggregate data on the quantum of content/accounts taken down/suspended for various categories of content violating their T&Cs, such as bullying and harassment, hate speech, spam, CSAM, etc.[13]

However, the effectiveness of such voluntary transparency reporting has been contested at best. It has become increasingly clear that information disclosures in themselves can do little unless such information is made available in a form that can hold decision-makers to account.[14] Platforms can potentially use voluntary transparency reporting to provide a "market-friendly" initiative to retain legitimacy

10 'Case Study #3: Transparency Reporting' (*New America*) <http://newamerica.org/in-depth/getting-internet-companies-do-right-thing/case-study-3-transparency-reporting/> accessed 16 December 2023.

11 ibid.

12 Robert Gorwa and Timothy Garton Ash, 'Democratic Transparency in the Platform Society', *Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge University Press 2020) <https://www.cambridge.org/core/books/social-media-and-democracy/democratic-transparency-in-the-platform-society/F4BC23D2109293FB4A8A6196F66D3E41> accessed 10 November 2023.

13 Spandana Singh and Leila Doty, 'The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules' (*New America*, 9 December 2021) <http://newamerica.org/oti/reports/transparency-report-tracking-tool/>.

14 Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973 <https://doi.org/10.1177/1461444816676645> accessed 28 February 2023.

while avoiding meaningful accountability.[15] Scholars have even classified this approach as "transparency washing".[16]

An important factor hindering the effectiveness of transparency reports is that platforms themselves determine which metrics to make public and when. They also decide the granularity of the information so provided.[17] The choice of which metric to disclose appears to be determined, to some extent, by platforms' need to manage public perception. For instance, Facebook's transparency reports state that the platform proactively removes 80-90% of hate speech, painting a rather positive picture of Facebook's moderation of hate speech,[18] something for which it has faced public criticism across the globe.[19] This proactive rate sheds light on the percentage of hate speech taken down by Facebook's automated systems and is defined by Facebook as the "percentage of violating content that we found before people reported it".[20] However, another metric, the takedown rate, which Facebook does not disclose in its transparency reports and which would paint a more holistic picture of Facebook's content moderation, was leaked in the Facebook files. As per leaked documents, Facebook "may action as little as 3-5% of the hate" on its services.[21]

Similarly, scholars have criticised the aggregate exposure rate released by platforms,[22] which typically measures the percentage of views on violating content as

15 Nicolas P Suzor, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation'.

16 Monika Zalnieriute, '"Transparency-Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism' (2021) <https://papers.ssrn.com/abstract=3805492> accessed 13 December 2023.

17 Spandana Singh and Leila Doty (n 13).

18 Noah Giansiracusa, 'How Facebook Hides How Terrible It Is With Hate Speech' *Wired*<https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/> accessed 5 December 2023.

19 Paul (n 3).

20 'Hate Speech | Transparency Centre' <https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/> accessed 16 May 2024.

21Giansiracusa (n 18).

22 Facebook's Prevalence Rate is defined as, "Prevalence considers all the views of content on Facebook or Instagram and measures the estimated percentage of those views that were of violating

a proportion of the total views on all content.[23] These provide no insights into the exposure of harmful content to vulnerable groups and the harm caused by such exposure.[24]

On the other hand, crucial metrics have been missing in the reports of most platforms. For instance, most platforms either provide no or insufficient information on the number of appeals to their content moderation decisions, as well as the percentage of successful appeals.[25] The Santa Clara Principles, a set of principles developed by a group of civil society organisations and endorsed by major tech corporations to enhance transparency in content moderation, highlight the importance of providing statistics on the percentage of successful appeals for content flagged by automated systems.[26] However, no platform provides such information, according to a study conducted by Urman et al. on the transparency reports of 10 major companies in 2021.[27] Similarly, stakeholders have highlighted the importance of disclosure of information on whether the actioned content was flagged by users, the platform's automated systems, or trusted flaggers (i.e. flagging information).[28]

---

content" and Youtube's Violative View Rate (VVR) "helps determine what percentage of views on YouTube comes from content that violates our policies." See 'Prevalence | Transparency Centre' <https://transparency.meta.com/en-gb/policies/improving/prevalence-metric/> accessed 16 May 2024; 'Building Greater Transparency and Accountability with the Violative View Rate' (*blog.youtube*) <https://blog.youtube/inside-youtube/building-greater-transparency-and-accountability/> accessed 16 May 2024.

23 See Giansiracusa (n 18); Anna-Sophie Harling, Declan Henesy and Eleanor Simmance, 'Transparency Reporting: The UK Regulatory Perspective' (2023) 1 Journal of Online Trust and Safety <https://www.tsjournal.org/index.php/jots/article/view/108> accessed 5 December 2023.

24 ibid.

25 Aleksandra Urman and Mykola Makhortykh, 'How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms' (2023) 47 Telecommunications Policy 102477 <https://www.sciencedirect.com/science/article/pii/S0308596122001793> accessed 14 June 2023.

26 'Santa Clara Principles on Transparency and Accountability in Content Moderation' (*Santa Clara Principles*) <https://santaclaraprinciples.org/images/santa-clara-OG.png> accessed 16 December 2023.

27 Urman and Makhortykh (n 25).

28 Suzor (n 15); 'Santa Clara Principles on Transparency and Accountability in Content Moderation' (n 26).

However, most platforms don't disclose disaggregated data about flagging information, broken down by categories of violating content.[29]

When it comes to reporting on state notices, platforms have generally been more forthcoming.[30] However, disaggregated data on which organ of the state (judiciary or executive) or state authorities (law enforcement agencies or ministries) is not available uniformly across countries.[31]

a. Provisions under the DSA

As per the DSA, transparency reporting obligations should be proportional to the societal impact of intermediaries,[32] and hence, an incremental set of reporting obligations are imposed on all intermediaries except micro, small and medium enterprises (MSMEs) that are not very large online platforms (VLOPs)/ very large online search engines (VLOSEs).

As per Article 15, all intermediaries (except MSMEs that are not VLOPs/VLOSEs) are required to publish yearly transparency reports on content moderation, containing information on:[33]

- Orders issued by the state,[34] classified by type of illegal content, issuing member state, time taken to acknowledge receipt and to take action.[35]
- Notices submitted under Article 16 (notice and action)[36] classified by type of illegal content, action taken, whether it was taken based on the law and/or

29 Urman and Makhortykh (n 25).

30 ibid.

31 ibid.

32 See recital 2 of the Draft Commission Implementing Regulation 2023 laying down templates concerning the transparency reporting obligations of providers of intermediary services and of providers of online platforms under Regulation (EU) 2022/2065 of the European Parliament and of the Council.

33 DSA 2022, art 15.

34 State Orders under Articles 9 and 10 of the DSA.

35 DSA 2022, art 15(1)(a).

T&Cs, and the median time for the said action. Information on the number of notices submitted by trusted flaggers and the number of notices processed by automated means is also to be included.[37]

- Action taken on the intermediary's own initiative classified by type of illegal content or the T&Cs violated, detection method and type of restriction imposed.[38] Here, intermediaries must also disclose the use of automated tools, and the training and assistance provided to persons in charge of content moderation.
- Internal complaint handling systems,[39] including the number of complaints received through the internal complaint handling systems in accordance with the platform T&Cs, as well as, the basis for those complaints, median time for decisions taken and the number of times decisions were reversed in the case of online platforms.[40]
- The use of automated means for content moderation,[41] including qualitative description, the precise purposes for which they are used, their accuracy and error rates and any safeguards in place.

Further, online platforms (except MSMEs that are not VLOPs or VLOSEs) must include information on disputes settled in out-of-court settlement bodies and on suspensions imposed for misuse (posting of manifestly illegal content or submission of manifestly unfounded notices or complaints).[42]

---

36 Individuals/entities can notify providers of hosting services (including platforms) about the presence of illegal content on their services (see Article 16).

37 DSA 2022, art 15(1)(b) lays down the reporting obligations for hosting services.

38 DSA 2022, art 15(1)(c).

39 DSA 2022, art 15(1)(d).

40 For online platforms under Article 20 provisions.

41 DSA 2022, art 15(1)(e).

42 DSA 2022, art 24.

VLOPs and VLOSEs have additional obligations with respect to transparency reporting by making it mandatory for them to publish reports every six months.[43] They also need to disclose information on the human resources dedicated to content moderation, broken down by each official language in the EU, [44] and the qualifications and the linguistic expertise of such persons, and the training and support given to them. Further, information on the accuracy of automated tools used for content moderation must be broken down by each official language in the EU. The transparency reports by VLOPs and VLOSEs must be published in at least one of the official languages of the member states.[45]

All online platforms and search engines must publish information on average monthly active subscribers every six months,[46] and VLOPs and VLOSEs must also include information on average monthly users in each state of the EU in their transparency report.[47]

Public reporting will provide information to the general public, users, researchers, and oversight bodies, offering a bird's eye view of how platforms perform content moderation. The aggregate statistics on appeals, flagging mechanisms, especially trusted flaggers and automated systems, are welcome inclusions in transparency reporting. Moreover, reporting all statistics broken down by category of illegal content or content in violation of the T&Cs of the service provider is certainly a step forward in making transparency reporting more granular. Further, the qualitative information on human moderators (see Section 6.4) and automated tools (see Section 6.3) have traditionally been absent from transparency reports and are welcome inclusions.

43 DSA 2022, art 42(1).

44 Including for compliance with obligations under Article 16 (notice and action), 20 (internal complaint-handling) and 22(trusted flaggers).

45 DSA 2022, art 42(1).

46 DSA 2022, art 24(2).

47 DSA 2022, art 42(3).

b. Standardisation and Harmonisation

In their analysis of voluntary transparency reports of major platforms, including YouTube, Facebook, Instagram, Reddit, Twitter, and TikTok, Singh and Doty found that there is a lack of standardisation across platforms.[48] Different platforms categorise violating content differently, and even when there are comparable categories across platforms, they may be defined differently in their T&Cs.[49] Further, there is little public information on the methods employed by platforms to calculate the number of content pieces or accounts violating their T&Cs in reporting.[50] For instance, the answer to how platforms classify and count if a piece of content violates multiple T&Cs might vary. The different metrics and methodologies used by platforms to report aggregate statistics limit any meaningful cross-platform comparison. The absence of public information on how these calculations are made and how they might have evolved over time further limits any temporal comparison of platform data. This also contributes to a limited understanding of how changes in platform policy or design impact content moderation and online speech.

However, the diversity of platforms, including the differences in T&Cs and content moderation systems, might render cross-platform comparisons inherently difficult.[51] Thus, while it is important to facilitate a minimum threshold of transparency and comparability across platforms, it is equally important to have flexibility in reporting metrics to safeguard the diversity of platforms and online communities.[52]

48 Spandana Singh and Leila Doty (n 13).

49 See Daphne Keller, 'Some Humility About Transparency' (19 March 2021) <https://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency> accessed 23 March 2022; CDT Europe, 'CDT Europe Contribution to European Commission Public Consultation: Draft Implementing Regulation Laying down Templates Concerning the Transparency Reporting Obligations under the Digital Services Act' (2024) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-/A>.

50 Keller and Leerssen (n 2).

51 Spandana Singh and Leila Doty (n 13); Harling, Henesy and Simmance (n 23); Daphne Keller (n 49).

52 Spandana Singh and Leila Doty (n 13).

The DSA recognises this and empowers the European Commission (EC) to adopt implementing acts to lay down templates concerning the form, content and other details of transparency reports.[53] In accordance with the above, the EC has conducted a public consultation for a Draft Implementing Regulation.[54]

The Draft Implementing Regulation[55] provides templates for transparency reporting with the objective of maintaining an adequate level of accountability through "comprehensive and comparable reporting".[56] It lays down the quantitative and qualitative templates[57] for reporting by all platforms[58] and provides detailed instructions to fill the templates and make the reports publicly available.[59] This also includes the format of transparency reports and the categories of content that are illegal or incompatible with the platform's T&Cs.[60] The regulation also lays down the reporting periods for platforms and VLOPs/VLOSEs.[61] It also establishes baseline

53 DSA 2022, arts. 15(3) and 24(6).

54'European Commission - Have Your Say' (European Commission - Have your say) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-_en> accessed 12 December 2023.

55 Draft Commission Implementing Regulation laying down templates concerning the transparency reporting obligations of providers of intermediary services and of providers of online platforms under Regulation (EU) 2022/2065 of the European Parliament and of the Council.

56 Draft Implementing Regulation 2023, recital 1.

57 Draft Implementing Regulation 2023, Annex 1.

58 Draft Commission Implementing Regulation laying down templates concerning the transparency reporting obligations of providers of intermediary services and of providers of online platforms under Regulation (EU) 2022/2065 of the European Parliament and of the Council art 1(1).

59 Draft Implementing Regulation 2023, art 1(2).

60 Draft Implementing Regulation 2023, Annex 2.

61 As per article 2(1) of the Draft Implementing Regulation 2023, "The reporting period for providers of intermediary services, of hosting services, and of online platforms shall be from 1 January until 31 December." and as per article 2(2), VLOPs and VLOSEs "shall publish their transparency reports every six months, covering respectively periods from 1 January until 30 June and 1 July until 31 December." Further, article 2(3) mandates that transparency reports be made "publicly available at the latest by two months from the date of the conclusion of each reporting period".

granularity of all information provided in the transparency reports by mandating that data be "broken down at a minimum by calendar months".[62]

Platforms are required to provide information for fifteen high-level categories of illegal (and incompatible) content, with each category containing up to seven sub-categories.[63] This categorisation is based on the statement of reasons provided by the platforms to the DSA Transparency Database maintained by the EC.[64] Platforms must avoid double counting and select the most relevant high-level category when multiple categories apply.[65]

This delineation of categories is significant given that platforms often group different categories of violations in reporting, while scholars have highlighted the need for more granular category-wise reporting.[66] However, this categorisation under the Draft Implementing Regulation raises several questions. Several submissions to the EC point to the unclear delineation between content that violates EU law or member state law and content that violates the platform's T&C.[67] Mozilla, in their submission to the EC, highlights the lack of clarity on how platforms must report content that

62 Draft Implementing Regulation 2023, art 2(4).

63 Part II of Annex 2 lays down the high-level categories which include "illegal or harmful speech", "negative effects on civic disclosure or elections", "non-consensual behaviour", "protection of minors", "risk of public security", and "self-harm" "violence", "data protection and privacy violations", "unsafe, non-compliant or prohibited products", "scams and/or frauds", "IP infringements" among others. Notably, category 15 in this classification comprises "content in violation of the platform's terms and conditions", with the subcategories "age-specific restrictions", "geographical requirements", "goods/services not permitted to be offered on the platform", "language requirements", "nudity", and "not captured by any other category's keyword".

64 EC, 'Digital Services Act Transparency Database' <https://transparency.dsa.ec.europa.eu/> accessed 16 December 2023.

65 Draft Implementing Regulation 2023, Annex 2, Part I, para 5.

66 Spandana Singh and Leila Doty (n 13).

67 Global Network Initiative, 'GNI Submission on Digital Services Act Transparency Reports Consultation' (2024) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-/F3451829_en>; CDT Europe (n 49).

violates their T&Cs but does not meet the threshold of being illegal.[68] For instance, there might be content that is hate speech as per the platforms' T&Cs, but it does not meet the legal threshold of being classified as unlawful speech in a jurisdiction. This conflation of illegal and legal but harmful content might lead to misrepresentation, especially when there is no scope for additional qualitative data providing contextual information to aggregate metrics.[69] Further, it has been pointed out that the illegal content categories do not cover the entire spectrum of illegal content defined across EU member states.[70]

There are also limitations to comparability that can be achieved through the categorisation and reporting framework in the draft templates. For instance, different platforms may use different methodologies for counting and reporting aggregate statistics for moderation,[71] and different platforms might define similar categories of incompatible content differently,[72] making cross-platform comparison meaningless unless additional qualitative information is provided or linked in the transparency reports.[73]

Even as high-level categorisation of illegal and incompatible content might be useful to ensure a minimum threshold of transparency and comparability across platforms, there are risks associated with prescriptive universal categorisation. These categories might indirectly influence how platforms define their T&Cs, which can have an impact on how online speech is governed even beyond the EU.[74] It might even lead to the homogenisation of T&Cs, which will negatively impact the entry of smaller and

68 Mozilla, 'Digital Services Act - Transparency Reports Rules and Templates | Feedback from Mozilla' (EC 2024).

69 Center for Studies on Freedom of Expression and Access to Information, 'CELE's Submission on the Draft Delegated Act on Transparency Reports (Detailed Rules and Templates) under the DSA' <https://www.palermo.edu/Archivos_content/2024/cele/paper-dsa/dsa.pdf>.

70 Global Network Initiative (n 67).

71 Center for Studies on Freedom of Expression and Access to Information (n 69).

72 CDT Europe (n 49).

73 See ibid; Center for Studies on Freedom of Expression and Access to Information (n 69); Global Network Initiative (n 67).

74 Global Network Initiative (n 67).

newer platforms, the diversity of platforms available to users, and may even disincentivise platforms to come up with new T&Cs beyond these mandated subcategories.[75] The standardisation of transparency reporting in this sense can have a negative impact on platform diversity.[76]

There are also free speech risks associated with states defining categories of illegal and incompatible content for reporting. Although the EC has used the statements of reasons from platforms to arrive at this categorisation. It is probable that other states may, while imposing similar obligations, lay down such categorisation unilaterally through executive orders.[77] Periodic reporting of numbers on the content taken down across state-defined categories of illegal or harmful speech can lead to platforms over-removing content across certain categories considered a priority by states. This can become a site for collateral censorship, especially when overbroad and vague speech harms are used by states to curtail dissenting speech online. For instance, concerns about the dangers of disinformation are rising globally.[78] This is accompanied by the introduction of legislation criminalising disinformation across several states like Singapore,[79] Sri Lanka,[80] Turkey,[81] Tunisia[82] raising concerns

75 See 'My Senate Testimony About Platform Transparency' <https://cyberlaw.stanford.edu/blog/2022/05/my-senate-testimony-about-platform-transparency> accessed 12 October 2023; Global Network Initiative (n 67); CDT Europe (n 49); Mozilla (n 68).

76 Harling, Henesy and Simmance (n 23).

77 See for instance, Rule 3(1)(b) of the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021 in India. Through this delegated legislation, platforms are directed to prohibit certain categories of harmful content in their Terms and Conditions, including misinformation and information that "threatens the unity, integrity, defence, security or sovereignty of India".

78 Henley (n 3).

79 Human Rights Watch, 'Singapore: "Fake News" Law Curtails Speech' (*Human Rights Watch*, 13 January 2021) <https://www.hrw.org/news/2021/01/13/singapore-fake-news-law-curtails-speech> accessed 24 September 2022.

80 Niresh Eliatamby, 'BASL Demands Withdrawal of Anti-Terrorism and Online Safety Bills' (23 September 2023) <https://english.newsfirst.lk/2023/9/23/basl-demands-withdrawal-of-anti-terrorism-and-online-safety-bills>.razil

81 Ruth Michaelson, 'Turkey: New "Disinformation" Law Could Jail Journalists for Three Years' *The Guardian* (13 October 2022) <https://www.theguardian.com/world/2022/oct/13/turkey-new-disinformation-law-could-jail-journalists-for-3-years> accessed 17 May 2024.

about free speech and state censorship[83]. In such a context, platforms may be incentivised or even coerced to register more takedowns under state-defined categories of disinformation. They may also be incentivised to align T&C definitions to legal definitions.

The discussion above underscores the importance of transparency in state-platform interactions. Several submissions to the EC have demanded more granular and nuanced information on state orders against illegal content and user information.[84] For instance, the templates in the draft delegated legislation[85] do not provide information on which state authority issues the notice,[86] the number of pieces of content or user accounts implicated in each order,[87] the type of action taken by platforms, or the option for platforms to provide qualitative reasoning for noncompliance with state orders.[88] GNI has pointed to the risks of reporting compliance with state orders in binary terms and recommended a category of "partial compliance" and a qualitative template for platforms to share information on the internal processes and principles they follow while responding to state orders, including any concerns they may wish to publicly share.[89]

---

82 Simon Speakman Cordall, 'Tunisia Anti-Fake News Law Criminalises Free Speech: Legal Group' (*Al Jazeera*) <https://www.aljazeera.com/news/2023/7/18/tunisia-anti-fake-news-law-criminalises-free-speech-legal-group> accessed 17 May 2024.

83 Amnesty International, 'A Human Rights Approach To Tackle Disinformation Submission to the Office of the High Commissioner for Human Rights 14 April 2022' (2022) <https://www.amnesty.org/en/wp-content/uploads/2022/04/IOR4054862022ENGLISH.pdf> accessed 17 May 2024.

84 Global Network Initiative (n 67); CDT Europe (n 49); Center for Studies on Freedom of Expression and Access to Information (n 69); Access Now, 'Submission to the Consultation on the Implementing Regulation (EU) .../... Laying down Templates Concerning the Transparency Reporting Obligations of Providers of Intermediary Services and of Providers of Online Platforms under the DSA' <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-/F3451797_en>.

85 Draft Commission Implementing Regulation 2023, Annex 1, Quantitative Template 1.2.1 and 1.2.2.

86 Global Network Initiative (n 67); Center for Studies on Freedom of Expression and Access to Information (n 69); Access Now (n 84).

87 Global Network Initiative (n 67).

88 Center for Studies on Freedom of Expression and Access to Information (n 69); CDT Europe (n 49).

89 Global Network Initiative (n 67).

The potential for negative externalities of information disclosure must also be carefully studied. For instance, reporting on the time taken to respond to state orders might indirectly pressurise platforms to arrive at decisions faster, which may impact the quality of the review conducted.[90] Several submissions call for more disaggregated information on notices submitted by trusted flaggers, given that state institutions, including law enforcement agencies, can be designated as flaggers and use this route instead of state orders under Articles 9 and 10.[91]

Thus, designing reporting templates needs to take into consideration several factors, including negative externalities. Regulators should also be cognizant of the risks of imposing disproportionate burdens on smaller and newer platforms.[92] Several submissions to the EC have also pointed to the disproportionate burden the detailed templates could impose on smaller platforms with limited resources, especially those that do not employ industrial-style commercial content moderation systems.[93] Wikimedia highlights the challenges of capturing their volunteer community-led moderation decisions in the reporting templates proposed by the EC.[94] Reporting obligations must be proportionate to platform size and flexible enough to accommodate the diversity of content moderation systems.[95] Certain baseline metrics can provide some form of comparability, and additional information can be provided by service-specific metrics[96]and relevant qualitative information.

---

90 David Nosák, 'The DSA Introduces Important Transparency Obligations for Digital Services, but Key Questions Remain' (*Center for Democracy and Technology*, 18 June 2021) <https://cdt.org/insights/the-dsa-introduces-important-transparency-obligations-for-digital-services-but-key-questions-remain/> accessed 6 March 2024; Global Network Initiative (n 67).

91 Global Network Initiative (n 67); CDT Europe (n 49); Center for Studies on Freedom of Expression and Access to Information (n 69); Access Now (n 84).

92 Harling, Henesy and Simmance (n 23).

93 Global Network Initiative (n 67); Mozilla (n 68).

94 Wikimedia Foundation, 'Digital Services Act: Consultation on the European Commission's Implementing Regulation and Template Comments from the Wikimedia Foundation' <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-/F3451860_en>.

95 Global Network Initiative (n 67); Mozilla (n 68).

96 Harling, Henesy and Simmance (n 23); Spandana Singh and Leila Doty (n 13).

Further, harmonisation of reporting between jurisdictions might also not always lead to accountability, given the political and socio-economic differences.[97] Different metrics can be meaningful in different jurisdictions.[98] As noted previously, in countries with weaker rule of law or which empower the executive to unilaterally send content takedown/blocking orders to platforms, transparency on state orders becomes critical.

c. Limitations of Transparency Reporting

Mandatory transparency reporting can be adopted by Global South countries as many major platforms already release some form of voluntary reports. In fact, many jurisdictions have started imposing such obligations.[99] This can push platforms based outside of the US to also publish transparency reports.[100]

Transparency reports should be published in languages other than English, which is not followed by many Big Tech companies.[101] Even the DSA only obligates VLOPs and VLOSEs to publish transparency reports in at least one of the official languages of the member states,[102] and this might hamper the accessibility of transparency reports by other platforms for various users and stakeholders.

Mandating transparency reports with granular quantitative and qualitative data, as done in the DSA, presents a great opportunity to build upon and enhance existing voluntary reporting. However, it must be reiterated that aggregate data in transparency reports only reflect platforms' decisions on content takedown and their own assessments of high-level indicators like the prevalence rate of harmful

97 Urman and Makhortykh (n 25).

98 ibid.

99 See The IT (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021 in India, PL2630 in Brazil.

100 Non-US based companies display a higher propensity to not publish transparency reports. See Urman and Makhortykh (n 25).

101 ibid.

102 DSA 2022, art 42(2).

content.[103] This provides no information on "why or how content moderation decisions were taken"[104] or how summary statistics were calculated.[105] Thus, such reporting in itself provides no way to evaluate the accuracy or quality of platform decisions underlying these statistics or to gauge the fairness and consistency in the enforcement of their T&Cs.[106] This does not provide meaningful accountability for platform moderation decisions and design.[107] Thus, complementary transparency measures like data access to researchers, third-party audits and self-risk assessments are key, and together, these enable a more holistic understanding of platforms' content moderation.

6.3. Automated content moderation tools

Platforms increasingly rely on automated tools as they perform moderation at an unprecedented scale, and public and regulatory pressure mounts to take down harmful content.[108] The reliance on such tools for moderation became even more pronounced during the pandemic.[109] Various platforms have deployed algorithmic tools for copyright violations, terrorism, violence, hate speech, CSAM, NCII, spam

103 Keller and Leerssen (n 2).

104 Svea Windwehr and Jillian C. York, 'Thank You For Your Transparency Report, Here's Everything That's Missing' (*Electronic Frontier Foundation*, 13 October 2020) <https://www.eff.org/deeplinks/2020/10/thank-you-your-transparency-report-heres-everything-thats-missing> accessed 12 December 2023.

105 See Center for Studies on Freedom of Expression and Access to Information (n 69).

106 Keller and Leerssen (n 2); Suzor (n 15).

107 Svea Windwehr and Jillian C. York (n 104); Keller and Leerssen (n 2); Suzor (n 15).

108 Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7 Big Data & Society 2053951720943234 <https://doi.org/10.1177/2053951720943234> accessed 8 November 2023; Spandana Singh, 'Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content' (2019) 22 New America 1; Keller (n 4); Gorwa, Binns and Katzenbach (n 5).

109 Louise Matsakis and Paris Martineau, 'Coronavirus Disrupts Social Media's First Line of Defense | WIRED' [2020] *Wired*<https://www.wired.com/story/coronavirus-social-media-automated-content-moderation/> accessed 14 December 2023; JC Magalhães and C Katzenbach, 'Coronavirus and the Frailness of Platform Governance | Internet Policy Review' (2020) 9 Internet Policy Review <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-68143-2>.

and bot detection.[110] The extent to which automated tools govern speech online can be gauged from the fact that Meta recently reported more than 90% of the content actioned was proactively detected by automated systems for most of its violation categories.[111] It further disclosed that most of the time, violating posts and accounts are removed automatically before being viewed by any user, and at other times, violating content is sent to review teams.[112] However, there is no transparency regarding what proportion of content removed through automated detection undergoes human review before being actioned. Additionally, it is unclear how platforms decide which content flagged by automated systems should be subject to human review.[113]

Algorithmic commercial content moderation comprises different kinds of systems. There are hash-matching systems like PhotoDNA[114] or predictive machine learning (ML) tools like Perspective API[115] used for the classification of content, and many complex systems could likely employ a combination of some degree of matching and classification.[116]

110 Gorwa, Binns and Katzenbach (n 5).

111 'How Technology Detects Violations' (*Meta Transparency Centre*, 18 October 2023) <https://transparency.fb.com/en-gb/enforcement/detecting-violations/technology-detects-violations/> accessed 14 December 2023.

112 ibid.

113 Similarly, YouTube has reported around 93% of the videos removed were flagged by automated systems with over 73% of these videos being removed before they had more than ten views. *See* 'YouTube Community Guidelines Enforcement (April 2023-June 2023)' (*Google Transparency Report*) <https://transparencyreport.google.com/youtube-policy/removals?hl=en> accessed 14 December 2023.

114 'PhotoDNA | Microsoft' <https://www.microsoft.com/en-us/photodna> accessed 17 May 2024.

115 'Perspective API' <https://www.perspectiveapi.com/> accessed 17 May 2024.

116 Gorwa, Binns and Katzenbach (n 5).

a. The urgent need for accountability

Both hash-matching and predictive ML systems come with their own set of unique risks and accountability requirements.[117] There is little public information on the extent to which platforms are using predictive systems to detect new forms of harmful speech versus them using some form of pattern/hash matching to remove newer instances or variants of already removed content.[118] Overall, algorithmic content moderation is "opaque, unaccountable and poorly understood".[119]

Systems that rely on identifying duplicates or variations of harmful content can pose risks and need transparency and accountability. Automated hash-matching cannot take into account the context surrounding a particular piece of content. For instance, it cannot detect when content containing terrorist propaganda or extremist violence is used for journalistic reporting or human rights abuse documentation.[120] Similarly, the opaque nature of what content finds its way into hash databases like the Global Internet Forum to Counter Terrorism (GIFCT) has raised several concerns.[121] For instance, it is not known whether content is subject to human review before being added to the database, or whether human review is conducted when new pieces of content are being matched with the content in the database or if automatic removal is the norm.[122]

Similarly, for predictive systems, there is little public information on the accuracy and reliability of automated tools and the accuracy of classifiers depends on the type of content they are trained on.[123] Certain types of content, like toxic speech or hate

117 Gillespie, 'Content Moderation, AI, and the Question of Scale' (n 108).

118 ibid.

119 Gorwa, Binns and Katzenbach (n 5).

120 Gillespie, 'Content Moderation, AI, and the Question of Scale' (n 108).

121 The Global Internet Forum to Counter Terrorism (GIFCT) was founded in 2017 by Facebook, Microsoft, YouTube and X (formerly Twitter) to prevent terrorists and violent extremists from exploiting digital platforms. See 'GIFCT | Global Internet Forum to Counter Terrorism' (GIFCT) <https://gifct.org/> accessed 15 December 2023.

122 Gorwa, Binns and Katzenbach (n 5).

123 Singh (n 108).

speech and extremist or terrorist content, are highly dependent on context and nuance, and their definition is inherently subjective, making them harder to detect.[124]

The lack of contextual understanding of the difference between satire, critique, and resignification also makes complete reliance on these automated systems without any form of human oversight and accountability mechanisms undesirable. Detection tools aimed at removing extremist content often remove content documenting human rights abuses and war crimes.[125] A research study found that Perspective found tweets by prominent drag queens in the US to have a higher level of toxicity than white nationalist speech because it failed to understand the resignifications of offensive words and mock impoliteness in LGBTQ speech.[126]

Further instances of bias similar to AI systems deployed outside content moderation have come to the fore. Internal documents revealed that 90 percent of hate speech that was taken down by Facebook's "race-blind" algorithms was speech directed at white people and men, while hate speech against marginalised Black communities was often left undetected.[127]This prompted Facebook to overhaul its algorithm in 2020 when internal researchers highlighted the need to take into account historical marginalisation.[128]The fact that none of these decisions were taken in an open and

---

124 See Khari Johnson, 'Zuckerberg: It's Easier to Detect a Nipple than Hate Speech with AI' (*VentureBeat*, 25 April 2018) <https://venturebeat.com/ai/zuckerberg-its-easier-to-detect-a-nipple-than-hate-speech-with-ai/> accessed 15 December 2023; Singh (n 108); Gorwa, Binns and Katzenbach (n 5); Svea Windwehr and Jillian C. York (n 104).

125 'Documentation of War Crimes Disappeared by Automated Tools' (*Electronic Frontier Foundation*, 20 May 2019) <https://www.eff.org/tossedout/documentation-war-crimes-disappeared-automated-tools> accessed 14 November 2023.

126 Thiago Dias Oliva, Dennys Marcelo Antonialli and Alessandra Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 Sexuality & Culture 700 <https://doi.org/10.1007/s12119-020-09790-w> accessed 23 March 2022.

127 Elizabeth Dwoskin, Nitasha Tiku, and Craig Timberg, 'Facebook's Race-Blind Practices around Hate Speech Came at the Expense of Black Users, New Documents Show' *Washington Post* (21 November 2021) <https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/> accessed 2 November 2023.

128 Dwoskin et al. note that the internal decision-making was fraught with opacity and leaks on the "worst of the worst project", an internal research project at Meta, revealing how researchers' recommendations for the algorithm to only remove hate speech targeted at 5 most vulnerable

accountable manner and details have only been revealed through whistleblower leaks highlights platforms' unilateral power on what speech is permissible on the internet. This raises important questions for Global South countries. Do platforms invest in similar internal research projects to understand historical marginalisation and power relations in different societies across South Asia, the Middle East, Latin America, South East Asia and Africa? Do platforms understand the racial, ethnic, caste, and religious power asymmetries that shape speech in these countries when training automated content moderation tools?

Content moderation decisions have real-world consequences. Inscrutable and unaccountable systems operate with impunity and entrench existing power relations at the cost of marginalised voices.[129] If content moderation occurs at scale with fully automated systems working in the background, they also obfuscate the political nature of content moderation decisions.[130] Gillespie notes, *"Machine learning techniques shift our understanding of societal phenomena: from instances among collectives, to patterns among populations".*[131]

The lack of adequate public information on the accuracy of such systems and the datasets they are trained on becomes even more acute when it comes to languages other than English.[132] Recently, multilingual large language models have been deployed by major platforms for content moderation, but they also face several challenges in moderating content in non-English languages, especially lower-

---

categories of users including, "those who are Black, Jewish, LGBTQ, Muslim or of multiple races" was rejected by executives for fear of conservative backlash. See Elizabeth Dwoskin, Nitasha Tiku and Heather Kelly, 'Facebook to Start Policing Anti-Black Hate Speech More Aggressively than Anti-White Comments, Documents Show' *Washington Post* (3 December 2020) <https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/> accessed 2 November 2023.

129 Roberts (n 6).

130 Gorwa, Binns and Katzenbach (n 5).

131 Gillespie, 'Content Moderation, AI, and the Question of Scale' (n 108).

132 Md Saroar Jahan and Mourad Oussalah, 'A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing' (2023) 546 Neurocomputing 126232 <https://www.sciencedirect.com/science/article/pii/S0925231223003557> accessed 15 December 2023.

resource languages.[133] This is compounded by the fact that platforms are not incentivised to invest in research and development for low-resource languages, especially in countries that are not lucrative markets for them.[134] Very little information is available on the efficiency of these automated tools in detecting context-heavy speech, like hate speech, across different languages and dialects. Transparency in how these systems perform across languages and dialects might be the much-needed first step to bring attention to these deficiencies and incentivise research, as failures in content moderation can have drastic consequences for local communities.[135]

Given the large-scale deployment of automated moderation systems, their huge impact on online speech and the opacity with which they operate, there is an urgent need for accountability.

b. Transparency Measures in the DSA

The DSA takes a giant leap forward for transparency in automated moderation and mandates the following:

- T&Cs published by intermediaries, including platforms, should disclose "information on the policies, procedures, measures and tools used for content moderation including algorithmic decision-making".[136]
- Periodic reporting by hosting services, including platforms, should disclose information on the number of notices[137] processed by using automated means.[138]

133 Gabriel Nicholas and Aliya Bhatia, 'Lost in Translation: Large Language Models in Non-English Content Analysis' (The Center for Democracy & Technology (CDT) 2023) <https://cdt.org/insights/lost-in-translation-large-language-models-i n-non-english-content-analysis/.>.

134 Nicholas and Bhatia (n 9).

135 Gabriel Nicholas, 'The Dire Defect of "Multilingual" AI Content Moderation' *Wired*<https://www.wired.com/story/content-moderation-language-artificial-intelligence/> accessed 2 September 2024.

136 DSA 2022, art 14(1).

- Periodic reporting by intermediaries, including platforms, should disclose information on the use of automated tools for content moderation at their own initiative for illegal content /violation of T&Cs.[139]
- Periodic reporting by intermediaries, including platforms, should disclose qualitative information on the use of automated means for content moderation, including the precise purposes for which they are used, their accuracy and error rates and any safeguards in place.[140]
- Periodic reporting by VLOPs and VLOSEs must further break down the qualitative information on automated tools by official languages of member states.[141]
- Hosting services, including platforms, to provide impacted users with information on the use of automated means in taking decisions where applicable. The statement of reason to the impacted user must include "information on whether the decision was taken in respect of content detected or identified using automated means".[142]

Submissions to the EC in response to the Draft Implementing Regulation on transparency reporting have recommended including additional information on error rates for automated systems to enable meaningful accountability. These include qualitative information on how such error rates are defined by different platforms,[143] input criteria for calculating error rates,[144]and the changes in accuracy over time and, across languages and content categories.[145]

---

137 Individuals/entities can notify providers of hosting services (including platforms) about the presence of illegal content on their services (see Article 16).

138 DSA 2022, art 15(1)(b).

139 DSA 2022, art 15(1)(c).

140 DSA 2022, art 15(1)(e).

141 DSA 2022, art 42(2)(c).

142 DSA 2022, art 17(3)(c).

143 Center for Studies on Freedom of Expression and Access to Information (n 69); CDT Europe (n 49).

144 Mozilla (n 68).

The DSA presents a good step forward, given the absolute lack of public information on the deployment and accuracy of these systems. However, it remains to be seen whether the depth of information that is provided on these tools in T&Cs and transparency reports will be sufficient, given that platforms often resort to intellectual property to not reveal any meaningful information.[146] For meaningful transparency, platforms must explain what kind of technology or inputs from automated systems are used at what points in the content moderation system.[147] Platforms must reveal the relationship between human reviewers and automatic review to understand if there are any checks and balances and oversight on automated systems.[148]

The provisions under the DSA will shed some light on systems that were largely opaque till now. Mandating such disclosures in transparency reporting and T&Cs will undoubtedly be a step forward for Global South jurisdictions. However, the importance of complementary transparency measures like data access for researchers, risk assessments and audits to make the qualitative and quantitative information in these disclosures meaningful cannot be emphasised enough.[149] Comparative data on the deployment and accuracy of automated systems across languages and categories of content will be an extremely valuable data point for Global South jurisdictions to seek accountability from platforms and potentially push for more investment in both automated systems and human reviewers in neglected languages and regions.

However, it is important to acknowledge that there might still be a long way forward, given that algorithmic accountability is not easy to achieve and these systems

145 Global Network Initiative (n 67).

146 Gorwa, Binns and Katzenbach (n 5).

147 Svea Windwehr and Jillian C. York (n 104).

148 ibid.

149 AlgorithmWatch, 'Feedback from AlgorithmWatch to the European Commission on Digital Services Act: Transparency Reports, Rules and Templates' <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-/F3451854_en>; CDT Europe (n 49).

represent dynamic and ever-changing socio-technical assemblages which confound explanations or disclosure-based transparency.[150]

6.4. Disclosure of Terms and Conditions (T&Cs) and Other Internal Policies

Public-facing guidelines inform users on what speech is permitted on the platforms, what is not permitted, and why. They articulate the principles of the platform and legitimise their right to govern online speech.[151] Platforms have in the past been hesitant to disclose the specific policies, procedures and practices for moderation.[152]This can be gauged from the fact that it is only in the face of significant public pressure post the 2016 US General Elections that platforms like Facebook published detailed "Community Standards" and transparency reports started including content taken down in violation of these standards in 2018.[153]

A. Disclosure of Terms and Conditions (T&Cs)

Gillespie[154] notes that public-facing guidelines by platforms often reflect an iterative process of creation and updation based on the experiences of moderating content over time and in response to public controversies over content takedowns. As a result, platforms often unilaterally update their T&Cs, often in response to extraordinary events like the pandemic or the 2016 US Presidential Elections.[155] This may be done

150 Ananny and Crawford (n 14).

151 Tarleton Gillespie, 'Community Guidelines, or the Sound of No', Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media (Yale University Press 2018).

152 Roberts (n 6).

153 Gorwa and Ash (n 12).

154 Gillespie, 'Community Guidelines, or the Sound of No' (n 151).

155 Magalhães and Katzenbach (n 109); Evelyn Douek, 'The Internet's Titans Make a Power Grab' *The Atlantic* (18 April 2020) <https://www.theatlantic.com/ideas/archive/2020/04/pandemic-facebook-and-twitter-grab-more-power/610213/> accessed 22 September 2022.

through blog posts, announcements or minor editorial changes in existing guidelines[156] making it hard to track the amendments and changes to the rules.[157]

The DSA mandates that all intermediaries, including online platforms, inform users of any restrictions on user content, including details on content moderation policies, procedures and tools (including algorithmic decision-making and human review) and details of internal complaint handling mechanisms in their T&Cs.[158]These should be laid out in "clear, plain, intelligible, user-friendly and unambiguous language, and shall be publicly available in an easily accessible and machine-readable format."[159] Further, when the service is directed at minors, the T&Cs must be laid out in a way that will be understandable for them.[160]

Additionally, platforms are obliged to clearly state in their T&Cs, the policy for determining service misuse based on posting illegal content or submitting unfounded notices or complaints, as well as the resulting penalty.[161]

VLOPs and VLOSEs must provide concise, easily accessible and machine-readable summaries of their T&Cs, including the redressal mechanisms.[162] All VLOPs/VLOSEs must publish the T&Cs in the official languages of the member states where services are offered.[163]

All intermediaries must communicate any significant changes in the T&Cs to users.[164] The EC maintains a Digital Services Terms and Conditions Database that tracks the

156 Gillespie, 'Community Guidelines, or the Sound of No' (n 151).

157 Svea Windwehr and Jillian C. York (n 104).

158 DSA 2022, art 14(1).

159 DSA 2022, art 14(1).

160 DSA 2022, art 14(3).

161 DSA 2022, art 23(4).

162 DSA 2022, art 14(5).

163 DSA 2022, art 14(6).

164 DSA 2022, art 14(2).

T&Cs of intermediaries, including platforms and highlights the changes and updates in them to help users keep track of the dynamic and evolving nature of the T&Cs.[165]

Mandating disclosure of T&Cs is an important transparency mechanism; some Global South jurisdictions have also drafted similar regulations.[166] However, the degree of specificity and detail with which platforms will publish their T&Cs in response to the DSA remains to be seen. This is because determining what level of transparency constitutes "clear and unambiguous" T&Cs is hard to define.[167] There are also trade-offs between comprehensive documentation and user-friendly readability that need to be carefully deliberated upon.

Mandating disclosure of T&Cs in clear and unambiguous language, as well as notifying users about updates, is a necessary first step in gaining transparency on platforms' content moderation systems. This, however, in itself is not sufficient to provide accountability on how such T&Cs are decided or implemented.

T&Cs, as they are generally articulated by platforms, provide no information on how platforms arrived at the values underpinning the guidelines.[168] There have been demands for platforms to reveal more about how these rules are developed, including the consultation and stakeholders that contributed to their framing.[169]

Further, a clear articulation of the T&Cs does not necessarily shed light on how content is practically moderated because definitions in T&Cs are inherently subjective, and the task of interpretation requires more detailed rules than the ones made public to users.[170] Whistleblower revelations have in the past shown how

165 EC, 'Digital Services Terms and Conditions Database' <https://platform-contracts.digital-strategy.ec.europa.eu/> accessed 16 December 2023.

166 In India, the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021 now mandate publication of "Terms of service, privacy policy, annual terms/policies reminders, and other agreements of intermediaries" must be made available to users in 22 official languages.

167 David Nosák (n 90).

168 Gillespie, 'Community Guidelines, or the Sound of No' (n 151).

169 Svea Windwehr and Jillian C. York (n 104).

170 Tarleton Gillespie, 'The Human Labor of Moderation', Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media (Yale University Press 2018).

Facebook's internal guidelines for human moderators are much more detailed and complex.[171] These internal rulebooks that govern how moderators make crucial decisions on what speech is permissible, often balancing competing freedoms within a limited time, are never open for public deliberation or oversight.[172]

b. Disclosure of information on human moderation

Platforms, in the past, have not revealed information on the composition of their moderation teams, the extent of dependence on outsourced labour, their working conditions and the training and support provided to them in the complicated and often traumatic work which underpins how we experience speech online.[173] This lack of transparency in the human moderation processes and teams also undermines confidence in platform decisions, and often, users suspect bias in moderation.[174]

Scholars have suggested that platforms provide more information on the demographic composition of content moderation teams, as well as, what internal guidelines, processes and support exist for human moderators to make contextually relevant, informed and consistent decisions at scale.[175] Further, disclosure of information on human moderators for each language and the training they receive will be useful. It will be equally important to have information on languages for which there are no human moderators who are native speakers situated in local context.[176]

Such qualitative information on the automated and human processes required for content moderation can help supplement the details in the T&Cs and aggregate

171 Nick Hopkins, 'Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence' *The Guardian* (21 May 2017) <https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence> accessed 16 December 2023.

172 Gillespie, 'The Human Labor of Moderation' (n 170).

173 ibid.

174 Suzor (n 15).

175 ibid.

176 Svea Windwehr and Jillian C. York (n 104).

statistics in transparency reporting.[177] Access to granular data for researchers is also vital to understand how content moderation decisions are actually made and to evaluate the merits of such decisions and the consistency of their enforcement at scale (see Chapter 5).[178]

The DSA mandates that transparency reporting for all intermediaries must include information on "the training and assistance provided to persons in charge of content moderation". [179] Further, transparency reporting for VLOPs/VLOSEs must additionally include the following information: (i) the human resources dedicated to content moderation, broken down by each official language in the EU, including for compliance with obligations under notice and action by users, internal complaint handling system and notices submitted by trusted flaggers;[180] (ii) "the qualifications and linguistic expertise of such persons, as well as the training and support given to them".[181]

These provisions can be beneficial for Global South countries as well. Internal leaks have revealed that platform profits often determine the quantum of their investment in content moderation resource allocation across jurisdictions. [182] Consequently, platforms don't employ sufficient human moderators who possess the knowledge of the local languages, dialects and socio-political contexts of many Global South countries. [183] This information on the composition of moderation teams can be

177 Suzor (n 15).

178 ibid.

179 DSA 2022, art 15(c).

180 DSA 2022, art 42(2)(a).

181 DSA 2022, art 42(2)(b).

182 Ben Gilbert, 'Facebook Ranks Countries into Tiers of Importance for Content Moderation, with Some Nations Getting Little to No Direct Oversight, Report Says' *Business Insider* (5 October 2021) <https://www.businessinsider.in/tech/news/facebook-ranks-countries-into-tiers-of-importance-for-content-moderation-with-some-nations-getting-little-to-no-direct-oversight-report-says/articleshow/87263447.cms> accessed 17 May 2023.

183 See De Gregorio and Stremlau (n 9); Marwa Fatafta, 'Facebook Is Bad at Moderating in English. In Arabic, It's a Disaster' (*Rest of World*, 18 November 2021) <https://restofworld.org/2021/facebook-is-bad-at-moderating-in-english-in-arabic-its-a-disaster/> accessed 3 September 2024; Steve Steckflow, 'Why Facebook Is Losing the War on Hate Speech in

helpful in holding platforms accountable. Reputational harm often drives platforms to make investments in non-priority low-income markets.[184] However, the effectiveness of such disclosure policies might rest on power dynamics between platforms and users. As Ananny and Crawford have cautioned,[185]*"transparency can reveal corruption and power asymmetries in ways intended to shame those responsible and compel them to action, but this assumes that those being shamed are vulnerable to public exposure."* Disclosure of the distribution of human moderators across languages, their qualifications, working conditions and information on their training to understand local contexts might not be sufficient in itself to effect changes in platforms' inequitable resource allocation.

6.5. Notice to Impacted Users

Notice to impacted users is a necessary component of due process and an important form of transparency.[186]The Santa Clara Principles outline that individuals must be notified about the specific content that was found (or alleged) to be in violation of T&Cs, the specific clause it violated, the method of detection of the said content, whether automated flagging was used, and in case of state orders, the relevant provision of the local law violated.[187]

However, platforms' notices to users have often failed to provide users with sufficient information. Suzor noted in their study that users impacted by content takedown or account suspension often expressed confusion on what content or action could have resulted in platform sanction.[188] Users often find the platform's explanation for sanctions insufficient and develop their own "vernacular explanations" on why their

---

Myanmar' *Reuters* (15 August 2018) <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> accessed 4 September 2024.

184 De Gregorio and Stremlau (n 9).

185Ananny and Crawford (n 14).

186 Suzor (n 15).

187 'Santa Clara Principles on Transparency and Accountability in Content Moderation' (n 26).

188 Suzor (n 15).

content was actioned.[189] Many users also express suspicions and concerns about being subjected to platforms' shadow banning or content demotions, which are typically done covertly by platforms without notifying users.[190]

The DSA mandates providers of all hosting services, including online platforms, to provide "clear and specific" statements of reasons to users who have been impacted by any restriction on the visibility of content (including removal, disabling access or demotion), demonetisation, suspension or termination of account for content that is found to be illegal or in violation of the T&Cs of platforms.[191] The DSA, in mandating that any form of visibility[192] or monetary restrictions[193] also be notified to users, effectively empowers users against secret shadow banning by platforms, which have raised grave concerns around opacity.[194] Notably, deceptive high-volume commercial content is exempt from such notification.[195] This might be necessary to prevent spammers from gaming the system against whom platforms have legitimate reasons to impose covert sanctions.[196]

---

189 ibid.

190 Gabriel Nicholas, 'Shedding Light on Shadowbanning' (Open Science Framework 2022) preprint <https://osf.io/xcz2t> accessed 11 September 2023.

191 DSA art 17(1) includes the following sanctions: "(a) any restrictions of the visibility of specific items of information provided by the recipient of the service, including removal of content, disabling access to content, or demoting content; (b)suspension, termination or other restriction of monetary payments;(c)suspension or termination of the provision of the service in whole or in part;(d)suspension or termination of the recipient of the service's account." Article 17(2) mandates that notification should apply when relevant electronic contact details of the users are known to the service providers.

192 One can get a better idea of what DSA implies by visibility restrictions on examining the Draft Implementing Regulation 2023 laying down templates concerning the transparency reporting obligations. Annex 2 includes, removal, disabling, demotion, age restriction, interaction restriction, labelling as subcategories of visibility restriction.

193 One can get a better idea of what DSA implies by monetary restrictions on examining the Draft Implementing Regulation 2023 laying down templates concerning the transparency reporting obligations. Annex 2 includes, suspension or termination of monetary payments.

194 Paddy Leerssen, 'An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation' (2023) 48 Computer Law & Security Review 105790 <https://www.sciencedirect.com/science/article/pii/S0267364923000018> accessed 18 July 2023.

195 DSA 2022, art 17(2).

196 Leerssen (n 194).

The DSA lays down that the statement of reasons must include: (i) information on the nature of the sanction and its territorial scope,[197] (ii) the facts and circumstances relied on for decision making,[198] and whether the decision was taken pursuant to a notice or platforms' own voluntary initiative,[199] (iii) whether automated means were used in the decision,[200] (iv) the legal provision or the T&C rule violated,[201] and (v) possibilities of redress available to the users.[202]

Similarly, notification of users in case of state orders for content removal is mandated. Intermediaries are required to inform impacted users about the state order and their action in response. This notification should include a statement of reasons,[203] the possibilities for redress and the territorial scope of the state order.[204]

Providing statements of reasons to explain decisions is essential to avoid arbitrariness and ensure fairness of application.[205] The DSA also emphasises that information provided under notice must "allow the recipient of the service concerned to effectively exercise the possibilities for redress".[206] It is important for users to understand how their content was identified for review, whether it was through automated detection, a complaint by other users, trusted flaggers or a state order. This information has typically not always been provided by platforms in the past.[207]

---

197 DSA 2022, art 17(3)(a).

198 DSA 2022, art 17(3)(b).

199 DSA 2022, art17(3)(b) also mandates that the statements of users contain the identity of the notifier where strictly necessary. There is no threshold defined for "strictly necessary" and leaving this to platform's discretion might endanger marginalised voices.

200 DSA 2022, art 17(3)(c).

201 DSA 2022, art 17(3)(d) and (e).

202 DSA 2022, art 17(3)(f).

203 As per DSA art 9(2)(a)(ii), statement of reasons in state orders must explain, "why the information is illegal content, by reference to one or more specific provisions of Union law or national law in compliance with Union law".

204 DSA 2022, art 9(5).

205 Suzor (n 15).

206 DSA 2022, art 17(4).

207 Suzor (n 15).

In addition to user notifications, the DSA also mandates online platforms to submit "statements of reasons" for their content moderation decisions to the EC to be included in a publicly accessible machine-readable database.[208] This public database aims to aid understanding of content moderation at a systems level, and the Transparency Database [209] so established has already started informing policy decisions. As discussed previously, it has been instrumental in shaping the templates under the Draft Implementing Regulation for transparency reporting.[210]

Complementary data access measures, as well as more qualitative information on content moderation processes, can further contextualise the information in the Transparency Database and enable more meaningful transparency.

Notifying users with a well-reasoned statement, including information on available redress mechanisms, will be a major step for users in the Global South to hold platforms accountable for their content moderation decisions. In order to make these notices more accessible, they must be in the language of the content of users and easily accessible.[211] Mandating such notifications for state orders will be critical to achieving transparency in the Global South. It will also be valuable to have a public database of anonymised statements of reasons for state requests of content removal/blocking.

208 DSA 2022, art 24(5).

209 EC (n 64).

210 Annex 2 of the draft implementing regulation 2023 lays down categories of illegal and incompatible content based on the DSA transparency database.

211 'Santa Clara Principles on Transparency and Accountability in Content Moderation' (n 26).

Insights for the Global South

- Mandatory Transparency Reporting can be a good start for the Global South. In order to make transparency more meaningful, diverse stakeholder consultations must drive discussions on the quantitative and qualitative information and the level of granularity to be outlined in legislation and delegated rules. Transparency on state orders, not only in terms of aggregate numbers, but also in terms of qualitative information on the platform's internal processes while responding to state orders, is critical.

- Standardisation of baseline metrics and templates for transparency reports across platforms has some merits. However, an overly prescriptive approach that relies on state classification of illegal and incompatible content can negatively impact small and medium platforms, innovation and diversity in T&Cs and moderation systems, and may even indirectly implicate the free speech of users. It is important that each jurisdiction carefully deliberate on the right balance. Harmonisation of reports for comparability and granularity should not undermine the diversity and innovation of platforms. Further, such standardisation must guard against states indirectly influencing platforms' T&Cs and content moderation decisions.

- Similarly, harmonising baseline transparency reporting across jurisdictions can also prove helpful in holding platforms accountable, provided such discussions have adequate representation from the Global South. However, more detailed metrics must be determined based on the local contexts of jurisdictions.

- Apart from aggregate numbers in reporting, qualitative information on the use of automated systems and the oversight mechanisms in place, as well as the training, support and qualifications of human moderators, is especially relevant for the Global South. The disclosure of the accuracy of automated tools broken down by language and category of content is crucial. Information on linguistic qualifications and demographic details of human moderators is critical to understanding the distribution of platform resources across regions

and communities. This information can be a starting point for advocating for major changes in the design and operation of platforms.

- The accessibility of T&Cs and transparency reports should be ensured by translating these into local languages and dialects.

- Notices to impacted users with detailed statements of reasons will help users exercise their rights and hold platforms and states accountable.

- Public databases on anonymised statements of reasons can help stakeholders, including researchers and policymakers, better understand how content moderation operates at a systemic level. Such databases can also contain state notices for content takedown.

- However, it is crucial to understand that mandating information disclosures in themselves cannot necessarily lead to greater accountability of platforms. The power dynamics between users and platforms, as well as platforms and states in the Global South, are important determinants. Further, transparency mechanisms that place responsibility on the users to interpret information and hold platforms accountable might be based on the presumption of an empowered audience, which might not always be the case for some Global South countries.

KEY INSIGHTS FOR THE GLOBAL SOUTH

In the preceding chapters of this report, [1] we have examined transparency mechanisms under the Digital Services Act (DSA) and assessed their suitability for Global South jurisdictions. We studied transparency mechanisms aimed at platforms' recommendations to users (see Chapter 1), advertisements hosted by platforms (see Chapter 2), systemic risk management by platforms (see Chapter 3), external audits to evaluate platforms' compliance (see Chapter 4), data access for public interest research (see Chapter 5), and content moderation practices (see Chapter 6).

In each chapter, we analysed the potential opportunities and challenges that each transparency mechanism presents for Global South jurisdictions, taking into account the socioeconomic, political and cultural contexts that may impact their adoption and operationalisation. We summarised key insights for the Global South towards the end of each chapter in the report. This does not necessarily imply that similar risks or challenges do not exist in the EU or other Global North countries. However, these may manifest differently in the Global South,

In this concluding chapter, we consolidate the "Key Insights for the Global South" across all the transparency mechanisms examined in the report for ease of reference. We recommend that the reader refer to the respective chapters for a detailed discussion. We acknowledge the limitations of evaluating the Global South as a single category, given the diversity of political and regulatory structures, economic conditions, and social and cultural factors constituting each jurisdiction and region. However, we hope these insights will serve as a starting point for future research into various aspects of platform transparency, with a focus on specific regions or jurisdictions in the Global South.

1 The DSA has been brought into force in a phased manner, and several delegated legislations associated with the transparency mechanisms outlined in this report are in different stages of deliberation, adoption and implementation. This report reflects the developments in legislation and implementation as of 31st October 2024.

Transparency In Recommending Content

- Users in the Global South exhibit a very limited understanding of the role of recommender systems in delivering content, and many perceive their content-feeds as neutral representations of reality. In this context, disclosure of the use of recommender systems and the parameters used by such systems, can offer users at least an elementary understanding as to why content appears to them in the given manner and priority.

- Nonetheless, parametric disclosures (as required by the DSA) cannot, by themselves, explain how recommender systems interact with the information ecosystems in which they operate. Such systems are fundamentally socio-technical in nature, and their outputs are shaped by platforms' design-choices, organisational processes as well as users' behaviour, alongside other key factors. Other (more systemic) transparency mechanisms, such as audits, risk assessments and researcher access to data, are expected to be instrumental in shedding more light on such factors.

- The DSA's user-facing disclosures assume an empowered user, capable of and willing to interpret such disclosures and make decisions accordingly. However, given the low-to-moderate levels of literacy, digital literacy and technical literacy currently prevailing in Global South jurisdictions, many users are particularly unlikely to be able to access and derive meaningful insights from such disclosures.

- Major platforms' advertising-based business models, based on optimising for users' engagement, are central to many risks that they create or propagate. In recent years, political pressure has forced certain platforms to incorporate other considerations, such as diversity and factual accuracy, in recommending content. Transparency requirements can bring platforms' priorities and choices into sharper focus. However, they must be complemented with measures in competition law to secure for users the effective choice to move to other platforms, if the content recommended by a platform does not align with their priorities and value-systems.

- The DSA requires VLOPs and VLOSEs to provide at least one option to users to access recommended content without profiling them. However, such optionality may not be very meaningful, particularly in the Global South, where most users may only have a basic understanding of how recommender systems operate and of the broader risks posed by profiling. Thus, platforms should not be permitted to profile a user to deliver recommended content, until and unless the user has expressly and affirmatively consented to it, after being informed of the associated risks in adequate detail, in a manner that is understandable and clearly accessible.

- The collection and use of personal data lie at the heart of personalised recommendations. Pertinently, DP laws provide controls and place safeguards on the collection and use of personal data by any entity, including platforms. While DP laws are being increasingly adopted since the introduction of the GDPR, many Global South jurisdictions are yet to enact such a law. Thus, Global South jurisdictions contemplating transparency frameworks for recommender systems must, on priority, adopt DP laws to undergird them.

Transparency in Advertising

- Advertisement transparency is crucial for Global South countries. There is an urgent need to study how microtargeting of ads plays out in postcolonial societies with multiple social cleavages and younger political systems. There has been little research or understanding of the discrimination and harms that such practices cause, both in terms of the distribution of economic opportunities as well as their impact on voter manipulation and offline violence.

- Ad transparency can help raise general awareness and understanding of how ads operate and empower citizens to engage with questions of privacy, discrimination and fair and democratic elections.

- Mandating advertisement repositories comprising both commercial and political ads with detailed information on sponsors, financial spending, and targeting methods employed by advertisers, including targeting parameters, will be an important step forward from voluntary ad archives for the Global South. It is important to note that this additional transparency obligation is only applicable to VLOPs and VLOSEs under the DSA, as this might be a resource-intensive obligation for smaller platforms.

- User-facing disclaimers provide baseline transparency to users and can be useful to Global South users as well. However, more research should be undertaken to understand the efficacy of such disclaimers in different social, cultural, and economic contexts to design effective disclaimers for users with differing levels of digital literacy.

- Similarly, providing an option for users to control targeting parameters appears to be a good step forward. However, the real accountability derived from such a measure must be critically examined, and more holistic methods to provide meaningful control which goes beyond abstract technical parameters should be studied.

- Transparency on targeting parameters is a good step forward, however, any meaningful accountability from platforms would also need information on how platforms classify users into interest groups for advertisers.

- Limited state regulatory and enforcement capacity to monitor and audit the adequacy of information disclosed through these transparency mechanisms can be a limitation in the Global South. Further, platforms often raise jurisdictional issues, making it difficult for regulators and civil society actors to hold them accountable in local courts.

Risk Management

- The risk management framework under the DSA exhibits a shift towards an *ex ante* systemic approach to intermediary regulation, where VLOPs and VLOSEs are required to pre-emptively and periodically assess the potential societal risks that may arise from the use of their services.

- The framework is expected to prompt platforms to consider their operational risks proactively and methodically, instead of reacting to harms as they magnify. If platforms were to similarly evaluate risks posed by their services in Global South jurisdictions, the insights gained could significantly assist regulators, public stakeholders, and platforms themselves, in formulating responses to address such risks.

- Specifying and encoding the categories of societal risks that platforms should assess can be a thorny task. Such risk-categories must be firmly tethered to values that have legal as well as normative acceptance, in the particular jurisdiction. Global South states that have not instituted a human rights framework in their constitutional documents or domestic law must contemplate alternative frameworks, to ground any risk management obligations that they seek to impose on platforms.

- The dynamic and contextual nature of societal risks posed by online platforms represents another challenge to the identification of risk-categories in law. On one hand, broad formulations of risk-categories may be difficult for platforms to assess, and would nudge them to restrict or demote borderline-legal content to comply with their obligations, particularly in the absence of the relevant cultural and linguistic expertise. On the other hand, highly prescriptive formulations can result in a framework that becomes anachronistic with changes in the socio-political context.

- Any Global South jurisdiction considering a risk assessment framework must promote rigorous research to identify the categories of risks that online platforms pose in that jurisdiction, how such risks evolve over time, and how

such risks intersect with social, political and economic structures and processes in that jurisdiction.

- In addition to the identification of risk-categories, it is important to develop suitable benchmarks for mitigation of risks. Such benchmarks should at least offer guidance regarding the threshold of risk at which mitigation measures should be implemented, and the threshold that such measures must meet to be considered adequate. However, it is important to understand that unlike outcome-based mechanisms (like notice-and-action), procedural frameworks like risk assessments are limited in their "enforceability" and cannot be tied to particular outcomes.

- Considering the complexity and the politically contentious nature of risk mitigation measures, a diverse range of stakeholders (including those affected disproportionately by the risks) must be meaningfully engaged in their formulation. Any Global South jurisdiction contemplating a risk management mechanism should consider making such engagement mandatory.

- Intermediaries must report the results of the risk assessments conducted and mitigation measures adopted, to the general public. Such reports would provide other intermediaries reference-points and encourage the development of industry best-practices and standards. To maximise the transparency gains from such reports, all redactions to public versions of risk assessment and mitigation reports must be grounded in principles of reasonability and proportionality.

Audits

- Regulatory audits, conducted by external auditors and overseen by regulatory authorities, can be an effective mechanism to systematically illuminate platforms' systems, policies and procedures. The information gathered through such audits can potentially assist stakeholders in understanding the propagation of information via platforms in the Global South, and in affixing

accountability on platforms for the adverse effects of their services in the region.

- Global South states must equip relevant regulatory bodies with adequate resources and independent powers to meaningfully process and critically assess audit reports under platform-audit frameworks, verify their contents and draw learnings that can inform the evolution of platform regulation.

- Clear benchmarks and methodologies are central to the reliability of audits, without which audits can be exploited by platforms to evade accountability. At the same time, auditing procedures must respect differences between platforms and the risks they pose in divergent contexts. Accordingly, Global South states should formulate benchmarks and methodologies tailored to their respective jurisdictional contexts as well as to differences between the risks posed by different kinds of services. As an initial step, they should build capacity to formulate such benchmarks and methodologies, and to contribute meaningfully in international initiatives, including multistakeholder forums and standard-setting bodies.

- Only very few organisations across the world currently possess the resources and expertise to conduct audits. Such limitations are particularly acute in the Global South. This heightens the risk of audit-capture by platforms, particularly if audits are commissioned by platforms themselves. Accordingly, Global South states should consider fostering an ecosystem of independent audits conducted by third parties acting in the public interest.

- In any case, given that issues relating to human rights and platform accountability have been extensively and predominantly examined by civil society organisations, independent researchers and other human rights practitioners, auditing frameworks in the Global South must provide pathways for the active engagement of such third parties in auditing as well the processes for formulation of auditing benchmarks and methodologies.

- Global South states must navigate limitations on their regulatory capacity and equip regulatory bodies with adequate financial and technical resources, and

independent powers to meaningfully assess audit reports and draw learnings that can inform the evolution of platform regulation.

- Towards maximising the transparency gains from audits, audit reports must be made public. While it may be necessary to redact certain information from audit reports, any redaction must be strictly proportional to the countervailing interest sought to be protected. Global South states should institute robust data protection legislation, to meaningfully balance transparency alongside privacy considerations.

Researcher Access to Platform Data

Data access for research is one of the most promising transparency mechanisms in the DSA. It will unlock platform data for independent public-interest expert scrutiny for the first time. This presents an immense opportunity to critically examine the online information ecosystem and platforms' moderation and curation of user-generated content and advertisements. As a result, data access for research is extremely valuable for countries located outside the EU too, including those in the Global South, where the information asymmetry is even more stark.

However, effectively operationalising complex data access mechanisms, like several other provisions in the DSA, requires strong societal structures, including communities of researchers, empowered civil society actors, and a favourable economic, political and regulatory environment to ensure free, independent and impactful research.[2]

Many countries across the Global South face several challenges to this effect:

- It appears that the power asymmetry between States and Big Tech, as well as Big Tech and Global South researchers, might hinder the ability of most Global South countries to mandate and operationalise such a complex and resource-intensive transparency mechanism in the near future.

2 See Martin Husovec, 'Will the DSA Work?' [2022] Verfassungsblog <https://verfassungsblog.de/dsa-money-effort/> accessed 14 August 2024.

- Researcher access to data, as mandated under Article 40(4), comes with a significant regulatory burden. This includes vetting researchers and research applications, as well as determining the modalities for meaningful data access. This requires independence, expertise, infrastructure, and technical, administrative and financial resources, which can prove to be a challenge for many regulators in the Global South at the moment.

- Data access for public interest research requires independent and bipartisan vetting of researchers. This could be a challenge in several Global South countries which do not have independent digital regulators and where the executive wields discretionary power to regulate platforms and online speech.

- In many Global South countries, there is a considerable risk that law enforcement agencies could gain access to APIs and tools intended for researchers or obtain the data collected by researchers. This poses serious concerns about privacy violations and the potential for increased surveillance.

- The declining academic freedom in several Global South countries can be a significant challenge in maintaining the independence of the research agenda and ensuring the safety of researchers. This also means that the scope of research must be carefully deliberated to prevent it from being dominated by state interests.

- The absence or inadequacy of data protection legislation in several Global South countries can impact both the privacy of users and the ability of researchers to gain access to platform data. It is essential to have privacy and data protection legislation with derogations for public-interest research and codes of conduct for ethical and privacy-protecting research practices.

- Inadequate funding and infrastructure for data processing, lack of data management and analysis skills, and insufficient institutional support in terms of ethics codes and data security codes might be challenging for several Global South researchers. Thus, allocating public funds for research and capacity building, as well as establishing institutional collaborations with Global North research organisations, could be beneficial.

- Several submissions to the EC have suggested that vetted researchers should not be restricted to those residing in the EU. Similarly, platforms must make APIs and tools under Article 40(12) available to researchers beyond the EU. This can pave the way for Global South researchers to collaborate with institutions and researchers in the EU to study platform data. However, Global South researchers are likely to face several barriers, including resource and funding constraints, as well as inter-jurisdictional legal conflicts limiting data transfers. It is also important that participation from Global South researchers in international collaborations must go beyond mere representation and be equal and meaningful for all researchers.

As a starting point, researchers in Global South can be provided with mandated access to (i) public data through API(s) and tools and (ii) legal immunity for independent data collection methods like data scraping for public interest research. Although these mechanisms also present challenges pertaining to data privacy and state surveillance, countries can aim to build robust legislation, safeguards, codes of practice and independent bodies. For the long-term, mandated researcher access, similar to that envisioned in Article 40(4), can be pursued. Starting with access to standardised datasets, this can progress to custom data demands as institutions mature and researchers gain more experience and skills.

Transparency in Content Moderation

- Mandatory Transparency Reporting can be a good start for the Global South. In order to make transparency more meaningful, diverse stakeholder consultations must drive discussions on the quantitative and qualitative information and the level of granularity to be outlined in legislation and delegated rules. Transparency on state orders, not only in terms of aggregate numbers, but also in terms of qualitative information on the platform's internal processes while responding to state orders, is critical.

- Standardisation of baseline metrics and templates for transparency reports across platforms has some merits. However, an overly prescriptive approach that relies on state classification of illegal and incompatible content can

negatively impact small and medium platforms, innovation and diversity in T&Cs and moderation systems, and may even indirectly implicate the free speech of users. It is important that each jurisdiction carefully deliberate on the right balance. Harmonisation of reports for comparability and granularity should not undermine the diversity and innovation of platforms. Further, such standardisation must guard against states indirectly influencing platforms' T&Cs and content moderation decisions.

- Similarly, harmonising baseline transparency reporting across jurisdictions can also prove helpful in holding platforms accountable, provided such discussions have adequate representation from the Global South. However, more detailed metrics must be determined based on the local contexts of jurisdictions.

- Apart from aggregate numbers in reporting, qualitative information on the use of automated systems and the oversight mechanisms in place, as well as the training, support and qualifications of human moderators, is especially relevant for the Global South. The disclosure of the accuracy of automated tools broken down by language and category of content is crucial. Information on linguistic qualifications and demographic details of human moderators is critical to understanding the distribution of platform resources across regions and communities. This information can be a starting point for advocating for major changes in the design and operation of platforms.

- The accessibility of T&Cs and transparency reports should be ensured by translating these into local languages and dialects.

- Notices to impacted users with detailed statements of reasons will help users exercise their rights and hold platforms and states accountable.

- Public databases on anonymised statements of reasons can help stakeholders, including researchers and policymakers, better understand how content moderation operates at a systemic level. Such databases can also contain state notices for content takedown.

- However, it is crucial to understand that mandating information disclosures in themselves cannot necessarily lead to greater accountability of platforms. The power dynamics between users and platforms, as well as platforms and states in the Global South, are important determinants. Further, transparency mechanisms that place responsibility on the users to interpret information and hold platforms accountable might be based on the presumption of an empowered audience, which might not always be the case for some Global South countries.

APPENDIX

(A downloadable version of the table in Excel format can be accessed here: DSA Summary Table)

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
				❖ **All intermediaries** ❖ **All intermediaries (except MSMEs that are not VLOPs/VLOSEs)** ❖ **All hosting service providers** ❖ **All online platforms** ❖ **All online search engines** ❖ **All online platforms (except MSMEs that are not VLOPs)** ❖ **All online platforms and all online search engines (except MSMEs that are not VLOPS/VLOSEs)** ❖ **All VLOPs and VLOSEs**	❖ **General public** ❖ **All users of the relevant service** ❖ **Affected users** ❖ **Regulatory authorities** ❖ **Independent experts, including vetted researchers, independent auditors and trusted flaggers**	❖ **Continuous** ❖ **Periodic (annually or semi-annually)** ❖ **Event-based** ❖ **Request-based**	❖ **Transparency in recommending** ❖ **Transparency in advertising** ❖ **Risk management** ❖ **Audits** ❖ **Researcher access to platform data** ❖ **Transparency in content moderation** ❖ **Disclosures to regulatory authorities (other than those covered under the mechanisms above)**
1	**Article 9(1)**	Orders to act against illegal content	Obligation to inform relevant state authorities about any action taken in response to state orders on illegal content.	All intermediaries	Regulatory authorities	Event-based	Disclosures to regulatory authorities
2	**Article 9(5) [Note: Read with Article 9(2).])**	Orders to act against illegal content	Obligation to notify the impacted user of the state order on illegal content and the consequent action taken. This should include: (i) a statement of reasons explaining why the content is illegal; (ii) redress options available to the user; and (iii) the territorial scope of the state order.	All intermediaries	Affected users	Event-based	Transparency in content moderation

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
3	**Article 10(1)**	Orders to provide information	Obligation to inform relevant state authorities about any action taken in response to state orders on access to user information.	All intermediaries	Regulatory authorities	Event-based	Disclosures to regulatory authorities
4	**Article 10(5) [Note: Read with Article 10(2).]**	Orders to provide information	Obligation to notify the impacted user of the state order on access to information and the effect given to it. This should include: (i) a statement of reasons explaining the objective for which the said information is required and why providing it is necessary and proportionate; and (ii) redress options available to the user.	All intermediaries	Affected users	Event-based	Transparency in content moderation
5	**Article 11**	Points of contact for Member States' authorities, the Commission and the Board	Obligation to designate a single point of contact for communication with the Member State authorities, the EC and the Board.	All intermediaries	Regulatory authorities	Continuous	Disclosures to regulatory authorities
6	**Article 12**	Points of contact for recipients of the service	Obligation to designate a single point of contact for users and make this information easily accessible and public.	All intermediaries	All users of the relevant service	Continuous	Transparency in content moderation
7	**Article 13(4)**	Legal representatives	Obligation to designate a legal representative (in one of the Member States) if the intermediary is not based in the EU but provides services there.	All intermediaries (which do not have an establishment in the Union but which offer services in the Union)	Regulatory authorities	Continuous	Disclosures to regulatory authorities

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
8	**Article 14(1), (2) and (3)**	Terms and conditions	Obligation to inform users of any restrictions on user content, including: (i) details on content moderation policies, procedures and tools (including algorithmic decision-making and human review) and (ii) details of internal complaint handling mechanisms in their T&Cs. Further, users must be informed of any significant changes to the T&Cs.	All intermediaries	General public	Continuous	Transparency in content moderation
9	**Article 14(5) and (6)**	Terms and conditions	Obligation to publish the T&Cs in the official languages of the Member States where services are offered and provide a machine-readable summary of the T&Cs including available remedies and redress mechanisms.	All VLOPs and VLOSEs	General public	Continuous	Transparency in content moderation
10	**Article 15(1)(a)**	Transparency reporting obligations for providers of intermediary services	Obligation to publish transparency reports on content moderation including information on state orders classified by: (i) type of illegal content; (ii) the issuing Member State; (iii) the median time taken to acknowledge the receipt and to take action pursuant to the order.	All intermediaries (except MSMEs that are not VLOPs/VLOSEs)	General public	Periodic (annually)	Transparency in content moderation

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
11	**Article 15(1)(b)**	Transparency reporting obligations for providers of intermediary services	Obligation to publish transparency reports including information on content moderation pursuant to notices submitted under Art 16 classified by the type of alleged illegal content, action taken and whether it was taken on the basis of the law or T&Cs and the median time for the said action. Information on the number of notices submitted by trusted flaggers and the number of notices processed by automated means should also be provided.	All hosting service providers (except MSMEs that are not VLOPs/VLOSEs)	General public	Periodic (annually)	Transparency in content moderation
12	**Article 15(1)(c) [Note: Read with Article 15(3).]**	Transparency reporting obligations for providers of intermediary services	Obligation to publish transparency reports including information on content moderation at the intermediary's own initiative. This should include information on the use of automated tools, training for human reviewers and the various restrictions imposed on users including measures taken to affect the availability, visibility and accessibility of user content. This information should be classified according to the: (i) category of illegal content or the T&C violated; (ii) the detection method; and (iii) the type of restriction imposed.	All intermediaries (except MSMEs that are not VLOPs/VLOSEs)	General public	Periodic (annually)	Transparency in content moderation

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
13	**Article 15(1)(d) [Note: Read with Article 15(3).]**	Transparency reporting obligations for providers of intermediary services	Obligation to publish transparency reports on content moderation including information on the complaints received through the internal complaint-handling systems. This should include information on the decisions made, the average time it took to make those decisions and the number of decisions reversed.	All intermediaries (except MSMEs that are not VLOPs/VLOSEs)	General public	Periodic (annually)	Transparency in content moderation
14	**Article 15(1)(e)**	Transparency reporting obligations for providers of intermediary services	Obligation to publish transparency reports including information on the use of automated means for content moderation, including (i) a qualitative description; (ii) the precise purposes for which they are used; (iii) indicators on accuracy and error rates; and (iv) the safeguards in place.	All intermediaries (except MSMEs that are not VLOPs/VLOSEs)	General public	Periodic (annually)	Transparency in content moderation
15	**Article 16(1)**	Notice and action mechanisms	Obligation to have accessible mechanisms facilitating submission of notices against illegal content.	All hosting service providers	General public	Continuous	Transparency in content moderation

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
16	**Article 16(4), (5) and (6)**	Notice and action mechanisms	Obligation to notify the users including (i) acknowledging receipt of the notice, (ii) the decision made, (iii) the redress mechanisms available, and (iv) whether or not automated means were used to process or make decisions regarding their notice.	All hosting service providers	Affected users	Event-based	Transparency in content moderation
17	**Article 17**	Statement of reasons	Obligation to notify the impacted users with a statement of reasons for any adverse action taken for illegal content or content in violation of T&Cs upon user notice or the service provider's own voluntary initiative (except pursuant to an order under Article 9). The statement of reasons should include information on: (i) the adverse action taken; (ii) the facts and circumstances of the case; (iii) the grounds for violation; (iii) the use of automated means; and (iv) redressal mechanisms available.	All hosting service providers	Affected users	Event-based	Transparency in content moderation
18	**Article 20(1), (2) and (3)**	Internal complaint-handling system	Obligation to establish a free and accessible internal complaint-handling system facilitating users to lodge complaints against decisions taken by platforms on the grounds of content being illegal or violating the platform T&Cs.	All online platforms (except MSMEs that are not VLOPs)	Affected users	Continuous	Transparency in content moderation

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
19	**Article 20(5)**	Internal complaint-handling system	Obligation to inform complainants of the decisions made under the platform's internal complaint-handling system and the available redressal mechanisms including out-of-court dispute settlement.	All online platforms (except MSMEs that are not VLOPs)	Affected users	Event-based	Transparency in content moderation
20	**Article 21(1)**	Out-of-court dispute settlement	Obligation to provide users with accessible information on out-of-court dispute settlement mechanisms.	All online platforms (except MSMEs that are not VLOPs)	Affected users	Continuous	Transparency in content moderation
21	**Article 23(4)**	Measures and protection against misuse	Obligation to clearly state in the terms and conditions the policy for determining service misuse based on posting illegal content or submitting unfounded notices or complaints, as well as, the resulting penalty of suspension.	All online platforms (except MSMEs that are not VLOPs)	General public	Continuous	Transparency in content moderation
22	**Article 24(1)(a) [Note: Read with Article 24(6).]**	Transparency reporting obligations for providers of online platforms	Obligation to additionally include the following information, in the report under Article 15: (i) number of disputes submitted to the out-of-court dispute settlement bodies; (ii) outcomes of the dispute settlement; (iii) median time needed for completing the dispute settlement procedures; and (iv) share of disputes where the online platform implemented the decision of the body.	All online platforms (except MSMEs that are not VLOPs)	General public	Periodic (annually)	Transparency in content moderation

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
23	**Article 24(1)(b) [Note: Read with Article 24(6).]**	Transparency reporting obligations for providers of online platforms	Obligation to additionally include the following information, in the report under Article 15: (i) total number of suspensions imposed pursuant to Article 23; and (ii) respective numbers of suspensions enacted for provision of illegal content, submission of unfounded notices and submission of unfounded complaints.	All online platforms (except MSMEs that are not VLOPs)	General public	Periodic (annually)	Other transparency mechanisms for content moderation
24	**Article 24(2)**	Transparency reporting obligations for providers of online platforms	Obligation to publish information on the average monthly active users of the service, calculated as an average over the past 6 months.	All online platforms and all online search engines (except MSMEs that are not VLOPS/VLOSEs	General public	Periodic (semi-annually)	Transparency in content moderation
25	**Article 24(3)**	Transparency reporting obligations for providers of online platforms	Obligation to provide the updated number of active users of the service (as per Article 24(2)), to state authorities, along with additional information regarding the calculation, including explanations and substantiation in respect of the data used.	All online platforms and all online search engines	Regulatory authorities	Event-based	Administrative disclosures to regulatory authorities
26	**Article 24(5)**	Transparency reporting obligations for providers of online platforms	Obligation to submit the decisions and the statements of reasons under Article 17(1) for inclusion in a publicly accessible and machine-readable database managed by the EC.	All online platforms (except MSMEs that are not VLOPs)	General public	Event-based	Transparency in content moderation

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
27	**Article 26(1)**	Advertising on online platforms	Obligation to ensure that users are able to identify clearly, concisely, unambiguously and in real-time: (i) that the information is an advertisement, including through prominent markings; (ii) the natural or legal person on whose behalf the information is presented; (iii) the natural of legal person who paid for the advertisement; and (iv) meaningful information directly and easily accessible about the main parameters used to determine the users to whom the advertisement is presented and where applicable, how to change those parameters.	All online platforms (except MSMEs that are not VLOPs)	All users of the relevant service	Continuous	Transparency in advertising
28	**Article 26(2)**	Advertising on online platforms	Obligation to provide users of the service with a functionality to declare whether the content they provide is or contains commercial communications.	All online platforms (except MSMEs that are not VLOPs)	All users of the relevant service	Continuous	Transparency in advertising

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
29	**Article 27(1) and (2)**	Recommender system transparency	Obligation to set out in plain and intelligible language, in T&Cs, the following: (i) the main parameters used in recommender systems, explaining why certain information is suggested to the users, including (a) the most significant criteria in determining the information; and (b) the reasons for the relative importance of those parameters; and (ii) any options for users to modify or influence the parameters.	All online platforms (except MSMEs that are not VLOPs)	All users of the relevant service	Continuous	Transparency in recommending
30	**Article 27(3)**	Recommender system transparency	Obligation to provide a functionality (directly and easily accessible from the specific section of the online interface where the information is being prioritised) that allows users to select and to modify their preferred option at any time, where several options are available for users to modify or influence the main parameters of the recommender system.	All online platforms (except MSMEs that are not VLOPs)	All users of the relevant service	Continuous	Transparency in recommending
31	**Article 34 [Note: Read with Article 42(4)(a)]**	Risk assessment	Obligation to identify, analyse and assess any systemic risks stemming from their design or functioning of their services or their related systems (including algorithmic systems), or from their use: (i) dissemination of illegal content;	All VLOPs and VLOSEs	Regulatory authorities + General public [Note: Read with Article 42(4)(a)]	Risk assessment: Periodic (annually) + Continuous (prior to deploying any functionality that is likely to have a critical impact on the risks identified in Article 34(2)) Submission of supporting	Risk management

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
			(ii) any actual or foreseeable negative effects for the exercise of FRs, particularly, the FRs to human dignity, respect for private and family life, the protection of personal data, freedom of expression and information (including the freedom and pluralism of the media), non-discrimination, respect for the rights of the child and a high-level of consumer protection, (iii) any actual or foreseeable negative effects on civic discourse and electoral processes, and public security; and (iv) any actual or foreseeable negative effects on gender-based violence, protection of public health and minors and serious negative consequences to the person's physical and mental well-being. Obligation to take into account, when conducting risks assessments, whether and how the following factors influence the systemic risks referred to in Article 34(1): (i) design of recommender systems and any other relevant algorithmic systems; (ii) content moderation systems; (iii) applicable T&Cs and their enforcement; (iv) systems for selecting and			documents to state authorities: Request-based + Periodic (upon completion) Publication of a report setting out the results of the risk assessment: Periodic (three months after the receipt of the report for each audit, which is required to be conducted annually) [Note: Read with Article 42(4)(a)]	

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
			presenting advertisements; (v) data-related practices. Further, obligation to analyse whether and how the systemic risks are influenced by intentional manipulation of the service, including by inauthentic use or automated exploitation, as well as the amplification and potentially rapid dissemination of illegal content and information that is incompatible with T&Cs. Additionally, obligation to preserve the supporting documents of the risk assessments for at least 3 years and to communicate them to state authorities upon request.				
32	**Article 36(1)(c)**	Crisis response mechanism	Obligation to report to the EC the precise content, implementation and the impact (qualitative and quantitative) of the specific measures taken to prevent, eliminate or limit any contribution to the serious threat to public health or public security in the EU or any significant part of it, as identified under Article 36(1)(a).	All VLOPs and VLOSEs	Regulatory authorities	Event-based (by a certain date or at regular intervals, as may be specified in the EC's decision to institute a crisis response mechanism)	Disclosures to regulatory authorities

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
33	**Article 37(1) and (2) [Note: Read with Article 42(4)(c).]**	Independent audit	Obligation to be subject to an independent audit at their own expense to assess compliance with: (i) the due diligence obligations under Chapter III; and (ii) any commitments undertaken pursuant to the codes of conduct under Articles 45 and 46 and the crisis protocols under Article 48. Further, obligation to afford the auditing organisations the cooperation and assistance necessary to conduct the audits in an effective, efficient and timely manner, including by giving them access to all relevant data and premises and by answering oral or written questions.	All VLOPs and VLOSEs	Regulatory authorities + General public [Note: Read with Article 42(4)(c).]	Conduction of audits: Periodic (at least annually) Submission of audit reports to state authorities: Periodic (upon completion) Publication of audit reports: Periodic (three months after the receipt of the audit report) [Note: Read with Article 42(4)(c)]	Audits

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
34	**Article 37(4) and (5) [Note: Read with Article 42(4)(c).]**	Independent audit	Obligation to ensure that the auditor establishes a written, substantiated audit report for each audit, including at least, the following: (i) the time period covered; (ii) the co-ordinates of the VLOP/VLOSE and the auditing organisation; (iii) a declaration of interests; (iv) a description of the specific elements audited and the methodology applied; (v) a description and a summary of main findings; (vi) a list of third parties consulted in the audit; (vii) an audit opinion on compliance of the VLOP/VLOSE with obligations and commitments referred to in Article 37(1) ('positive', 'positive with comments' or 'negative'); (viii) where the audit opinion is not 'positive', operational recommendations on specific measures to achieve compliance and the recommended timeframe; and (ix) where the auditor is unable to audit specific elements or to express an audit opinion based on its investigations, an explanation of the circumstances and reasons why those elements could not be audited.	All VLOPs and VLOSEs	Regulatory authorities + General public [Note: Read with Article 42(4)(c).]	Conduction of audits: Periodic (at least annually) Submission of audit reports to state authorities: Periodic (upon completion) Publication of audit reports: Periodic (three months after the receipt of the audit report) [Note: Read with Article 42(4)(c)]	Audits

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
35	**Article 37(6) [Note: Read with Article 42(4)(d).]**	Independent audit	Obligation to (i) adopt an audit implementation report, setting out measures taken to implement operational recommendations made pursuant to an audit; and (ii) justify the reasons for not implementing any operational recommendations and setting out any alternative measures taken to address any instances of non-compliance identified.	All VLOPs and VLOSEs	Regulatory authorities + General public [Note: Read with Article 42(4)(d).]	Adoption of audit implementation reports: Event-based (within one month of receipt of recommendations under an audit report that is not 'positive') Submission of audit implementation reports to state authorities: Periodic (upon completion) Publication of audit implementation reports: Periodic (three months after the receipt of the audit report) [Note: Read with Article 42(4)(d).]	Audits

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
36	**Article 39(1)**	Additional online advertising transparency	Obligation to compile and make publicly available in a specific section of the online interface, through a searchable and reliable tool that allows multicriteria queries and through APIs, a repository containing the following information: (i) the content of the advertisement, including the name of the product, service or brand and the subject matter of the advertisement; (ii) the person on whose behalf it was presented and the person who paid for the advertisement; (iii) the period during which it was presented; (iv) whether the advertisement was intended to be presented to specific groups of users and if so, the main parameters used for the purpose (including the main parameters used to exclude one or more groups, if appliable); (v) the commercial communications published on the VLOPs and identified pursuant to Article 26(2); and (vi) the total number of users the service reached and the aggregate numbers for the group(s) of users (broken down by each state) that the advertisement specifically targeted.	All VLOPs and VLOSEs	General public	Continuous (for the entire period during which an advertisement is presented and until one year after the advertisement was presented for the last time)	Transparency in advertising

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
37	**Article 39(2)**	Additional online advertising transparency	Obligation to include, for each specific advertisment removed or made unavailable based on illegality or incompatibility with T&Cs, the following information in the repository: (i) the period during which it was presented; (ii) whether the advertisement was intended to be presented to specific groups of users and if so, the main parameters used for the purpose (including the main parameters used to exclude one or more groups, if appliable); (iii) the total number of users the service reached and the aggregate numbers for the group(s) of users (broken down by each state) that the advertisement specifically targeted; and (iv) a statement of reasons for removal of the advertisement, as per Article 17(3)(a) to (e) or, a reference to the legal basis for the removal, as per Article 9(2)(a)(i).	All VLOPs and VLOSEs	General public	Continuous (for the entire period during which an advertisement is presented and until one year after the advertisement was presented for the last time)	Transparency in advertising
38	**Article 40(1) and (2)**	Data access and scrutiny	Obligation to provide, within reasonable time, relevant state authorities access to data that is necessary to monitor and assess compliance with the DSA.	All VLOPs and VLOSEs	Regulatory authorities	Request-based	Disclosures to regulatory authorities
39	**Article 40(3)**	Data access and scrutiny	Obligation to explain the design, logic, functioning and testing of algorithmic systems, including recommender systems.	All VLOPs and VLOSEs	Regulatory authorities	Request-based	Disclosures to regulatory authorities

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
40	**Article 40(4), (5), (6) and (7)**	Data access and scrutiny	Obligation to provide access to data (through appropriate interfaces, including online databases or APIs) to vetted researchers for research that contributes to the detection, identification and understanding of systemic risks as set under Article 34(1) and, to the adequacy, efficiency and impact of risks mitigation measures pursuant to Article 35.	All VLOPs and VLOSEs	Independent experts (vetted researchers)	Request-based	Researcher access to platform data
41	**Article 40(12)**	Data access and scrutiny	Obligation to give access to data, including real-time data (where technically possible), provided that the data is publicly accessible in their online interface by researchers, including those affiliated to not-for-profit bodies, organisations and associations, who comply with the conditions set out in Article 40(8)(b) to (e), and who use the data solely for performing research that contributes to the detection, identification and understanding of systemic risks pursuant to Article 34(1).	All VLOPs and VLOSEs	Independent experts (researchers, including vetter researchers)	Continuous	Researcher access to platform data
42	**Article 41(4)**	Compliance function	Obligation to communicate the name and contact details of the head of the compliance function to relevant state authorities.	All VLOPs and VLOSEs	Regulatory authorities	Continuous	Disclosures to regulatory authorities

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
43	**Article 42(1) and (2)**	Transparency reporting obligations	Obligation to additionally include the following information, in the report under Article 15: (i) the human resources dedicated to content moderation, broken down by each official language in the EU, including for compliance with obligations under Article 16, 20 and 22; (ii) the qualifications and linguistic expertise of such persons, as well as the training and support given to them; and (iii) the indicators of accuracy of the automated tools used for content moderation and related information set out in Article 15(1)(e), broken down by each official language in the EU.	All VLOPs and VLOSEs	General public	Periodic (at least every six months)	Transparency in content moderation
44	**Article 42(3)**	Transparency reporting obligations	Obligation to additionally include the number of average monthly users of the service in each state of the EU, in the report under Article 15.	All VLOPs and VLOSEs	General public	Periodic (at least every six months)	Disclosures to regulatory authorities
45	**Article 42(4)(a) and (e)**	Transparency reporting obligations	Obligation to transmit to the relevant state authorities and make publicly available a report setting out the results of the risk assessment pursuant to Article 34.	All VLOPs and VLOSEs	Regulatory authorities + General public	State authorities or regulators: Periodic (upon completion) General public: Periodic (three months after the receipt of the report for each audit, which is required to be conducted annually)	Risk management

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
46	**Article 42(4)(b) and (e)**	Transparency reporting obligations	Obligation to transmit to the relevant state authorities and make publicly available the following: (i) the specific mitigation measures put in place pursuant to Article 35(1); and (ii) where applicable, information about the consultations conducted in support of risks assessments and design of the risk mitigation measures.	All VLOPs and VLOSEs	Regulatory authorities + General public	State authorities or regulators: Periodic (upon completion) General public: Periodic (three months after the receipt of the report for each audit, which is required to be conducted annually)	Risk management
47	**Article 42(4)(c)**	Transparency reporting obligations	Obligation to transmit to the relevant state authorities and make publicly available the audit report provided for in Article 37(4).	All VLOPs and VLOSEs	Regulatory authorities + General public	State authorities or regulators: Periodic (upon completion) General public: Periodic (three months after the receipt of the report for each audit, which is required to be conducted annually)	Audits
48	**Article 42(4)(d)**	Transparency reporting obligations	Obligation to transmit to the relevant state authorities and make publicly available the audit implementation report provided for in Article 37(6).	All VLOPs and VLOSEs	Regulatory authorities + General public	State authorities or regulators: Periodic (upon completion) General public: Periodic (three months after the receipt of the report for each audit, which is required to be conducted annually)	Audits

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
49	**Article 67**	Requests for information	Obligation to provide information to relevant state authorities to carry out their tasks under Section 4 (Supervision, investigation, enforcement and monitoring in respect of VLOPs and VLOSEs).	All VLOPs and VLOSEs	Regulatory authorities	Request-based (in relation to an investigation for a suspected infringement)	Disclosures to regulatory authorities
50	**Article 68**	Power to take interviews and statements	Power of the EC to interview any consenting person, with regard to an investigation of a suspected infringement.	All VLOPs and VLOSEs	Regulatory authorities	Request-based (in relation to an investigation for a suspected infringement)	Disclosures to regulatory authorities
51	**Article 69**	Power to conduct inspections	Powers of the EC to conduct inspections to carry out its tasks under Section 4 (Supervision, investigation, enforcement and monitoring in respect of VLOPs and VLOSEs), including: (i) to enter the premises, land and means of transport of the VLOP/VLOSE; (ii) to examine the books and other records, irrespective of the medium of storage; (iii) to require the VLOP/VLOSE or any concerned person to provide access to an explanation on its organisation, functioning, IT systems, algorithms, data-handling and business practices; and (iv) to take assistance of auditors or experts appointed by any relevant state authorities.	All VLOPs and VLOSEs	Regulatory authorities + Independent experts (including auditors)	Event-based (upon the discretion of state authorities)	Disclosures to regulatory authorities

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
52	**Article 72**	Monitoring actions	Power of the EC to (i) monitor the effective implementation and compliance with the DSA and to impose an order to retain all documents deemed to be necessary for the purpose; (ii) to order VLOP/VLOSE's to provide access to, and explanations relating to, its databases and algorithms; and (iii) to appoint independent external experts and auditors to assist the EC and to provide specific expertise or knowledge.	All VLOPs and VLOSEs	Regulatory authorities	Event-based (upon the discretion of state authorities)	Disclosures to regulatory authorities
66	**Article 75(2)**	Enhanced supervision of remedies to address infringements of obligations laid down in Section 5 of Chapter III	Obligation to communicate an action plan setting out the necessary measures sufficient to terminate or remedy an infringement, pursuant to a decision under Article 73, including the following: (i) commitment to perform an independent audit in accordance with Article 37(3) and (4) on the implementation of other measures; and (ii) commitment to participate, where appropriate, in a relevant code of conduct provided for in Article 45.	All VLOPs and VLOSEs	Regulatory authorities	Event-based (upon the finding of an infringement by state authorities)	Disclosures to regulatory authorities

Sl. No.	Article	Title of the provision	Summary of the obligation	Applicability to category of intermediaries	Information Recipient	Frequency of delivery of the information	Transparency mechanism
67	**Article 75(3)**	Enhanced supervision of remedies to address infringements of obligations laid down in Section 5 of Chapter III	Obligation to communicate the audit report under Article 75(2) to the EC and to keep the EC up to date on steps taken to implement the action plan.	All VLOPs and VLOSEs	Regulatory authorities	Event-based (upon the finding of an infringement by state authorities)	Disclosures to regulatory authorities

Centre for Communication Governance

National Law University Delhi

Sector 14, Dwarka

New Delhi – 110078

011- 28031265

ccgdelhi.org

privacylibrary.ccgnlud.org

twitter.com/CCGNLUD

ccg@nludelhi.ac.in