



Isaura Salgado Silva

The Governance of Expression Under Algorithmic Logic: A Roadmap to Safeguard Public Discourse from Commodification

Dissertation to obtain a Master's Degree in
Law, in the specialty of Public Law and Regulation

Supervisor:

Luís Heleno Terrinha, Professor at NOVA School of Law

September 2025



The Governance of Expression Under Algorithmic Logic:
A Roadmap to Safeguard Public Discourse from Commodification

September 2025

Anti-Plagiarism Statement

I, Isaura Salgado da Silva, hereby declare that this dissertation, titled 'The Governance of Expression Under Algorithmic Logic: A Roadmap to Safeguard Public Discourse from Commodification' is my original work and has not been submitted in part or in full for any degree or academic award in any university or educational institution. All sources used in this dissertation have been duly acknowledged and cited in the references section. I acknowledge that plagiarism is a serious academic offense, and I have taken all necessary measures to ensure that my dissertation is free of any plagiarism. In addition, I have used the appropriate citation methods and followed ethical academic practices in researching and writing this dissertation. I also understand that if any form of plagiarism is detected in my dissertation, it will result in severe consequences, including but not limited to the revocation of my degree, legal action, and the publication of my name and offense in relevant academic publications. By signing this anti-plagiarism statement, I affirm that I have taken full responsibility for the originality and integrity of my dissertation and have not engaged in any form of academic misconduct.

Lisbon, 15 September 2025



Assinatura Recuperável

X

Isaura Salgado Silva

Isaura Salgado Silva

Assinado por: trust_161de40e-d19b-4948-8c42-7e90dd660f8e

Acknowledgements

I would like to express my sincere gratitude to the faculty and staff of Nova, whose lessons and insights over the past two years have greatly shaped and enriched this work.

Above all, I am deeply thankful to my supervisor, Professor Luís Heleno Terrinha, for his generosity in accepting to supervise this dissertation. His insightful recommendations, constructive feedback, and continuous guidance were instrumental to this research.

On a personal note, I wish to thank my family—especially my parents, Graça and Eduardo—for their unwavering encouragement and support throughout my education and career. Their love has been a constant source of inspiration and motivation, without which this journey would not have been possible.

Formal Declarations

This dissertation is written in English and consistently follows UK spelling, grammar, and formatting conventions.

The layout is Calibri, size 12, and the line spacing is 1.5. The body occupies a total of 168 463 characters, including spaces and notes.

All citations are provided in accordance with the Oxford University Standard for the Citation of Legal Authorities (OSCOLA), using the notes and bibliography system.

The main specifications are as follows:

1. Books (in footnotes and bibliography, respectively):

Author, *Title* (Edition, Publisher Year) Page.

Author Surname Initial(s), *Title* (Edition, Publisher Year)

2. Journal Articles (in footnotes and bibliography, respectively):

Author, 'Title' (Year) Volume Journal Name First Page, Pinpoint Page.

Author Surname Initial(s), 'Title' (Year) Volume Journal Name First Page

3. Journalistic Articles (in footnotes and bibliography, respectively):

Author, 'Title' (*Publication Name*, Date) <URL> accessed Date.

Author Surname Initial(s), 'Title' (*Publication Name*, Date) <URL> accessed Date

4. Reports (in footnotes and bibliography, respectively):

Author/Organization, *Report Title* (Publisher, Date) Pinpoint Page <URL> accessed Date.

Author/Organization, *Report Title* (Publisher, Date) <URL> accessed Date

5. Websites & Blogs (in footnotes and bibliography, respectively):

Author/Organization, 'Title of Page or Post' (*Website Name*, Date) <URL> accessed Date.

Author/Organization, 'Title of Page or Post' (*Website Name*,
Date) <URL> accessed Date

Abstract

Initially celebrated as enablers of democratic expression, social media platforms have paradoxically become powerful gatekeepers, exerting unprecedented forms of control over speech that would have been unimaginable in the pre-digital era. Amid rising concerns over their role on social fragmentation—from youth mental health crises to electoral instability—regulators have begun scrutinizing these mechanisms of speech (content) control.

Early attempts at “new school speech regulation” focused on content removal, overlooking a far more structural and consequential form of speech governance. Platforms today shape public discourse not just through reactive takedown mechanisms, but *proactively*—via algorithms that sort, rank, and deliver content to users—impacting both commercial and non-commercial expression, regardless of illegalities or perceived harmfulness. While newer regulations have begun to address this reality—for example, the Digital Services Act introduces a framework aimed at mitigating systemic risks related to AI-driven “recommender systems”—responses remain overly reliant on procedural transparency and industry self-assessments, an approach that falls short of ensuring meaningful platform accountability.

Missing from the current regulatory landscape is a framework that directly confronts the structural role of algorithms in shaping speech circulation, prioritizes robust protections against the commercial manipulation of free expression and democratic participation, and explores more effective mechanisms of platform accountability to ensure genuine democratic oversight. This paper critiques gaps in existing literature, identifies key shortcomings in EU legislation on AI and platform accountability and calls for an urgent reorientation of regulatory debates towards these objectives. It advances both concrete measures to protect free speech and democratic participation online, and a theoretical foundation to guide such reforms in a way that aligns technological governance with democratic values.

Resumo

Inicialmente celebradas como facilitadoras da expressão democrática, as plataformas de mídia social tornaram-se paradoxalmente poderosos guardiões, exercendo formas

de controle inéditas sobre a expressão que seriam inimagináveis na era pré-digital. Em meio a crescentes preocupações sobre o seu papel na fragmentação social—desde crises de saúde mental entre os jovens até instabilidades eleitorais—reguladores têm voltado atenção a estes mecanismos de controle.

As primeiras tentativas de regulação *new school* de expressão concentraram-se na remoção de conteúdo, negligenciando uma forma de governança da expressão muito mais estrutural e consequente. Atualmente, as plataformas moldam o discurso público não apenas por meio de mecanismos reativos de remoção, mas de forma *proativa*—através de algoritmos que classificam, hierarquizam e distribuem conteúdo aos utilizadores—afetando tanto a expressão comercial quanto a não comercial, independentemente de ilegalidades ou suposta nocividade. Embora regulamentos mais recentes tenham começado a abordar essa realidade—por exemplo, o Regulamento dos Serviços Digitais (*Digital Services Act*) introduz um regime destinado a mitigar riscos sistêmicos relacionados aos “sistemas de recomendação” impulsionados por IA—suas respostas continuam excessivamente dependentes de transparência processual e autoavaliações do setor, uma abordagem insuficiente para garantir uma responsabilização significativa das plataformas.

O que falta no cenário regulatório atual é uma estrutura que enfrente diretamente o papel estrutural dos algoritmos na moldagem da circulação da expressão, que priorize proteções robustas contra a manipulação comercial da liberdade de expressão e da participação democrática, e que explore mecanismos mais eficazes de responsabilização das plataformas para garantir uma supervisão democrática genuína. Este trabalho critica lacunas na literatura existente, identifica deficiências-chave na legislação da União Europeia sobre IA e responsabilização de plataformas, e clama por uma reorientação urgente dos debates regulatórios para esses objetivos. Ele avança tanto medidas concretas para proteger a liberdade de expressão e a participação democrática online, quanto uma base teórica para guiar tais reformas de forma a alinhar a governança tecnológica com os valores democráticos.

Sumário

Abstract	vi
Resumo.....	vi
Introduction	2
The Rise of Algorithmic Gatekeeping: From the Marketplace of Ideas to Market-Driven Engagement	2
Methodology and structural overview	6
Aims and objectives	7
Premisses and limitations.....	8
Analysis	9
Part I - Algorithmic Control and the Structural Commodification of Speech	9
Chapter 1 – Why and How Algorithms Influence What We See	9
Chapter 2 –Public Discourse and the Normative Challenges of Algorithmic Content Distribution	13
Chapter 3 – A Brief Review of Literature on ‘New School’ Speech Regulation and its Gaps	16
Part II - A Critique of EU Regulation	20
Preliminary Considerations: A Brief Justification	20
Chapter 1 - The AI Act: An Attempt to Interpret Potential Impacts Amidst Legislative Silence.....	22
Chapter 2 - The DSA’s Attempt to Regulate Structural Controls With Superficial Procedural Rules	26
Chapter 3 - The DMA: A Missed Opportunity to Reshape the Market Dynamics that Underline Content Distribution ..	37
Chapter 4 - From Failure to Insight: What Can We Learn from Current Regulatory Standards	40
Part III: The Solution: A Roadmap for Intentional Reform	52
Chapter 1 – What is Free Speech and How to Protect it in the Digital Age: Constructing a Theoretical Foundation for Reform	52
Chapter 2 - Strengthening Balkin’s Framework Through Interdisciplinary Dialogue	59
Chapter 3 – Examining Concrete Regulatory Measures for Structural Reform	63
Concluding Remarks	71
Bibliography	74
Cases	74
EU Legislation & Official Documents	74
Other Legislation & Official Documents	74
List of References	75

Introduction

The Rise of Algorithmic Gatekeeping: From the Marketplace of Ideas to Market-Driven Engagement

In 1919, Justice Oliver Wendell Holmes introduced the ‘marketplace of ideas’ metaphor. In his dissent in *Abrams v. United States*, he argued that truth and consensus arise from the free competition of ideas, famously stating that “the best test of truth is the power of the thought to get itself accepted in the competition of the market.”¹ The concept of a self-correcting public sphere where expression circulates freely to foster shared understanding became an enduring universal reference point, influencing legal and philosophical doctrines throughout the twentieth century, in the United States and beyond.²

In Holmes’ model, the proper response to falsehood is not censorship or regulation, but more speech, allowing ideas to be challenged and corrected through open dialogue.³ This *laissez-faire* approach rests on the assumption of a neutral public square, a level playing field where all speakers enjoy equal opportunity to challenge opposing viewpoints and compete fairly.⁴ While this ideal has always been more aspirational than realistic—with access to resources and institutional power having historically influenced the visibility of perspectives, even in traditional media environments—⁵ it faces renewed challenges in the digital age.

When social media platforms emerged in the early 2000’s, they were ostensibly celebrated as enablers of democratic expression. By enabling decentralized, unfiltered, many-to-many communication, these platforms provided an apparently unmediated open environment that promised to dismantle the traditional barriers to democratic

¹ ‘The Constitution: Supreme Court Case Library: *Abrams v United States*’ (National Constitution Center) <https://constitutioncenter.org/the-constitution/supreme-court-case-library/abrams-v-united-states> accessed 21 February 2025

² The European Court of Human Rights (ECtHR) has frequently invoked the marketplace of ideas in its rulings, emphasizing the importance of pluralism and the free flow of information in democratic societies. For example, see: *Handyside v United Kingdom* (1976) 1 EHRR 737, [49].

³ Kyle Langvardt, ‘A New Deal for the Online Public Sphere’ (2019) 26 *George Mason Law Review* 1, 2 <<https://ssrn.com/abstract=3149513>> accessed 11 August 2025.

⁴ Ibid 9.

⁵ Ibid.

participation. To many, this environment resonated with Holmes' vision of an ideal marketplace where diverse voices could freely interact, challenge one another, and generate consensus through dialogue. No longer was public discourse mediated exclusively by a handful of privileged institutions broadcasting to passive audiences. Ordinary individuals were suddenly empowered to engage in public discourse and cultural production, taking on an unprecedented dual role as both their consumers and producers.⁶ This participatory architecture suggested nothing short of a revolution in democratic expression, theoretically dispersing power to allow every individual equal capacity to shape the collective narratives through which social reality is constructed.

Decades into this promised digital utopia, we are confronted with an unsettling paradox. While social media has indeed opened new avenues for participation, contrary to what Holmes expected, the proliferation of speech has not guided society towards truth or consensus.⁷ Instead, misinformation and polarization have become entrenched features of modern life.⁸ Ironically, nowhere are the signs of fragmentation stronger than in the very digital spaces that once promised democratic renewal.

The creeping realization that social media has failed its original emancipatory promise grew alongside a clearer understanding of its operational logic: a troubling interplay of surveillance, data extraction, behavioural manipulation and profit maximization.⁹ At the core of this system are data-driven algorithms that control and direct the delivery of published content to users—determining who gets to see to see what on social media. Beneath the veneer of personalization, these mechanisms not only monitor and predict

⁶ Jack M. Balkin, 'Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society' (2004) 79 *New York University Law Review* 1, 32. <<https://jackbalkin.yale.edu/sites/default/files/files/digitalspeechanddemocraticculture1.pdf>> accessed 11 August 2025.

⁷ "Yet it is far from clear that Americans' addiction to expressive online 'engagement' has produced a richer or wiser public discourse. To the contrary, it increasingly appears that social media products are eroding the mechanisms that allow the freedom of speech to police itself and to produce public goods." In Langvardt (n 3) 25.

⁸ Kai Riemer and Sandra Peter, 'Algorithmic Audiencing: Why we Need to Rethink Free Speech on Social Media' (2021) *Journal of Information Technology* 409, 413. <<http://journals.sagepub.com/doi/10.1177/02683962211013358>> accessed 13 August 2025.

⁹ In her work on informational capitalism, Julie E. Cohen describes how decentralizing cultural and political production has inadvertently fuelled "the emergence of dominant information platforms (...) affording new vantage points for data harvesting, surplus extraction, and manipulation of information flows." In Julie E. Cohen, *Between Truth and Power the Legal Constructions of Informational* (Oxford University Press 2019) 251 <<https://doi.org/10.1093/oso/9780190246693.001.0001>> accessed 13 August 2025.

user behaviour but also seek to actively skew it into generating more interactions, platform use and monetization opportunities.¹⁰

The influence exerted by these algorithms extends far beyond the manipulation of individual behaviour. By dictating the terms of content distribution, platforms gain control over media markets themselves.¹¹ For instance, Facebook's decision to prioritize original videos over other forms of content, and X's decision to deprioritize posts with external links, have both been linked to the reduction of referral traffic to traditional media outlets,¹² exemplifying the structural influence these internal commercial decisions can have on media economy. More critically, the ability to determine which manner of expression goes viral gives platforms a new form of control over cultural flows.¹³ By deciding the visibility, scale, and prominence of people and ideas across the virtual public sphere, these algorithms actively influence their perceived legitimacy and relevance, allowing platforms to shape public discourse at its source.¹⁴

Such powers have never been available to governments, at least not with the level of precision and depth made possible by recent advances in Artificial Intelligence (AI).¹⁵ As public discourse increasingly concentrates on social media, online visibility emerges as a key currency in attention economies, influencing social and political agendas in real life. While they can enable rapid social mobilization (e.g., the global spread of the Black Lives Matter movement in 2020),¹⁶ the architectures of online visibility can also entrench harmful dynamics and segregate users into polarized echo chambers,¹⁷ amplifying extremist ideologies (e.g., as seen in Facebook's admitted role in amplifying

¹⁰ Balkin (n 6) 21; Langvardt (n 3) 19; Riemer & Peter (n 8) 412.

¹¹ Langvardt (n 3) 4.

¹² Oliver Darcy, 'Publishers See Dramatic Drop in Facebook Referral Traffic as the Social Platform Signals Exit from News Business' *CNN* (17 August 2023) <<https://edition.cnn.com/2023/08/17/media/facebook-referral-traffic-reliable-sources>> accessed 15 August 2025; Danny D'Cruze, 'Elon Musk Confirms That X Deprioritises Posts With Links To Discourage "Lazy Linking"' *Business Today* (26 November 2024) <<https://www.bustoday.in/technology/news/story/elon-musk-confirms-that-x-deprioritises-posts-with-links-to-discourage-lazy-linking-455115-2024-11-26>> accessed 15 August 2025.

¹³ Langvardt (n 3) 4.

¹⁴ Ibid 5.

¹⁵ Balkin (n 6) 5.

¹⁶ Ho-Chung H. Chang and others, '#JusticeforGeorgeFloyd: How Instagram Facilitated the 2020 Black Lives Matter Protests' (07 December 2022) *PLOS ONE* <<https://doi.org/10.1371/journal.pone.0277864>> accessed 19 August 2025.

¹⁷ Nathan Matias and Lucas Wright, 'Impact Assessment of Human-Algorithm Feedback Loops.' (1 of March 2022) *Just Tech* <<https://doi.org/10.35650/JT.3028.d.2022>> accessed 14 August 2025.

anti-Rohingya violence in Myanmar).¹⁸ By unilaterally controlling the design of these algorithms, platforms yield unchecked power over the machinery of mass mobilization, giving rise to concerns over concentration of influence, erosion of democratic oversight, and the manipulation of public opinion.¹⁹

Amidst growing global anxiety over platforms' role in escalating societal harms—including misinformation,²⁰ racism,²¹ hate speech,²² political polarization,²³ electoral destabilization,^{24/25} public health degradation²⁶ and overall social fragmentation—the European Union (EU) enacted landmark digital legislation: the Digital Services Act (DSA) and the Digital Markets Act (DMA) in 2022, followed by the AI Act in 2024. While early regulations like the E-Commerce Directive were not equipped to respond to the emergence of algorithmic surveillance and predictive targeting in content delivery, the new regulatory triad was specifically designed to curb platform dominance and discipline AI deployment, raising hopes that social media platforms might finally face long-overdue oversight and accountability.

¹⁸ Amnesty International, *Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations – New Report* (29 September 2022)

<<https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/>> accessed 14 August 2025.

¹⁹ Jack M. Balkin, 'Free Speech is a Triangle' (2018) Yale Law School Public Law Research Paper No 640 2011, 2047 <<https://ssrn.com/abstract=3186205>> accessed 15 August 2025.

²⁰ Soroush Vosoughi, Deb Roy, and Sinan Aral, 'The Spread of True and False News Online' (2018) 359(6380) *Science* 1146, 1146 <<https://www.science.org/doi/10.1126/science.aap9559>> accessed 12 August 2025.

²¹ Eugenia Siapera and Paloma Viejo-Otero, 'Governing Hate: Facebook and Digital Racism.' (2021) 22(2), *Television & New Media* 112-130 <<https://doi.org/10.1177/1527476420982232>> accessed 22 July 2025.

²² Steve Stecklow, 'Why Facebook is Losing the War on Hate Speech In Myanmar' (15 August 2018) *Reuters Investigates* <<https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>> accessed 22 July 2025.

²³ Jaeho Cho and others 'Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization.' (2020) 64(2) *Journal of Broadcasting & Electronic Media*, 150, 172. <<https://doi.org/10.1080/08838151.2020.1757365>> accessed 22 July 2025.

²⁴ Ferenc Huszár and others 'Algorithmic Amplification of Politics on Twitter' (2022) 119(1) *Proceedings of the National Academy of Sciences* e2025334119 <<https://doi.org/10.1073/pnas.2025334119>> accessed 15 August 2025.

²⁵ Victor do Nascimento Silva and Regina Helena Alves Silva, 'Are Algorithms Affecting the Democracy in Brazil?' (International Symposium on Ethical Algorithms, December 2018) <https://www.researchgate.net/profile/Regina-Alves-Da-Silva/publication/332798661_Are_Algorithms_Affecting_the_Democracy_in_Brazil/links/5cca355d4585156cd7c1ac22/Are-Algorithms-Affecting-the-Democracy-in-Brazil.pdf> accessed 22 July 2025.

²⁶ Vincenzo Carrieri and others, 'Vaccine Hesitancy and (Fake) News: Quasi-Experimental Evidence from Italy' (2019) *Health Economics Letter* <<https://doi.org/10.1002/hecl.3937>> accessed 22 July 2025.

However, while it sometimes acknowledges the systemic threats of algorithmic speech governance (e.g., Recital 81 of the DSA highlights that algorithmic design can undermine free expression and information) the new framework has ultimately proven incapable of addressing the profound democratic deficits generated by platform capitalism in content distribution. The AI Act sidesteps the issue entirely; the DSA attempts to confront structural control with rules favouring superficial procedural compliance; and the DMA's regulation of algorithmic rankings focuses solely on the economic effects of discrimination, overlooking its broader impacts on fundamental free speech rights and public discourse. Collectively, these frameworks fall short of delivering a coherent, unified strategy to align algorithmic governance with values of democratic participation and free speech.

A critical normative challenge remains unresolved: how to curb platform control over public discourse and safeguard free speech from the distortive commodification pressures embedded in social media's content distribution systems. This study takes on that challenge, offering a dual contribution: first, by establishing a theoretical foundation for intentional regulation, and second, by proposing concrete measures for its implementation. Our intention is not to deliver a closed formula for regulating platform power, but to articulate a normative vision for a more democratic digital public sphere; capable of resisting market-driven distortions while fostering pluralism, democratic participation, cultural engagement, and innovation.

Methodology and structural overview

This study employs a qualitative and theoretical methodology, grounded in EU legal analysis and enriched by interdisciplinary research spanning law, communication, computer science, and political studies. Analysis unfolds in three sections:

- 1) Part I - Algorithmic Control and the Structural Commodification of Speech,** outlines the general mechanics of the algorithms that shape what we see on social media, and explains why and how they have become the driving force of platform business model. This section also identifies the democratic tensions inherent to an automated and privatized model of speech governance based on predictive patterning and outlines the normative challenges that guide

subsequent analysis. Finally, Part I delivers a brief literature review to expose the limitations of scholarship on ‘new school’ speech regulation.

- 2) Subsequently, **Part II - A Critique of EU Regulation**, offers a critical assessment of pertinent EU frameworks—the DSA, the DMA and the AI Act—highlighting their limitations in addressing the systemic threats of structural corporate control over speech. After each framework is analysed individually, this section distils four key lessons as to why they collectively fall short of promoting effective democratic oversight and meaningful platform accountability, providing a foundation for essential future reform.
- 3) Building upon the critique developed in the previous section, **Part III - A Roadmap for Intentional Reform**, constructs an actionable framework aimed at reconciling platform power with the participatory values of free speech. We begin by reconceptualizing free speech as an infrastructure for democratic participation and outlining a dynamic vision of protection centred on its essential participatory value. Grounded on this theoretical foundation and interdisciplinary scholarship, we derive concrete regulatory measures to safeguard and promote this vision.

Following this three-phase analysis, we offer concluding remarks, recapitulating the study’s scope, findings, limitations and significance.

Aims and objectives

Using the outlined methodology, this study seeks to contribute to the fields of free speech, democratic participation, and AI and platform regulation. The key contributions we aim to advance include:

- a) Raising academic awareness of the commodification pressures inherent to platform business models, identifying them as a structural threat to free speech and democratic discourse on social media.
- b) Demonstrating the limitations of current EU regulatory frameworks in addressing the systemic risks posed by these models to democratic communication.

- c) Establishing the theoretical foundations for regulatory reform, centred on safeguarding democratic discourse and promoting freedom of expression as a form of public infrastructure, beyond the traditional focus on individual rights.
- d) Proposing a roadmap that combines regulatory methodology with concrete policy measures to mitigate the commodification logic embedded in platform design, with the aim of realigning speech governance with democratic values.

Premises and limitations

This study does not presume to offer a definitive interpretation of recent frameworks like the DSA, the DMA and the AI Act. Instead, our intention is to highlight some of the prominent features of the current regulatory paradigm, expose the myths and fallacies that support them and challenge these assumptions, laying the foundation for a new normative vision in platform and AI regulation. We do not presume to provide a definite solution to the normative challenge identified, instead, this study navigates the options that open outside the limitations of current paradigms.

Throughout this study, we will periodically reference potential platform-related harms—such as misinformation propagation, political polarization, hate speech amplification, user fragmentation, and electoral integrity risks—for theoretical and illustrative purposes. These references should not be understood as asserting that algorithmic systems are the sole or primary causal factor in any specific instance of harm. We fully acknowledge the complex interplay of social, psychological, and technological dynamics underlying such harms,²⁷ and we make no claim to evidentiary certainty in attributing causation in particular cases. Rather, these harms are invoked as plausible consequences that flow from the core premises of this investigation, namely that:

²⁷ On the complexities of human-algorithm feedback loops, see: Matias and Wright (n 17). Suggesting that a well-functioning public sphere will not be significantly affected by algorithmic design, see: Benkler, Yochai and others, 'Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics' (2018) Oxford University Press <<https://library.oapen.org/viewer/web/viewer.html?file=/bitstream/handle/20.500.12657/28351/97801>90923624.pdf?sequence=1&isAllowed=y>> accessed 22 July 2025.

- a) Platforms possess both the structural capacity and economic incentives to influence what is seen and heard on social media.
- b) Because they can, platforms will actively and systematically amplify speech and discourse that advances their commercial interests, while simultaneously suppressing less commercially valuable speech.
- c) This practice inherently undermines fundamental free speech rights and public interest goals—namely, democratic participation.

Despite irreducible complexities—particularly regarding the ontological challenge of isolating algorithmic effects from human behaviour—our investigative focus remains valuable, and the methodology employed to pursue it is rigorous and sound.

Analysis

Part I - Algorithmic Control and the Structural Commodification of Speech

Chapter 1 – Why and How Algorithms Influence What We See

Social media platforms have always relied on automated systems to organize and deliver content to their users. In early years, these were relatively simple mechanisms, typically programmed to arrange posts in reverse chronological order.²⁸ However, as Jack M. Balkin observes, platforms want users “to experience digital products in ways that will encourage consumption and increase profits, and they want to structure the digital environment accordingly.”²⁹ To this end, from the mid-2000s onwards, chronological displays were abandoned in favour of more sophisticated systems that integrated surveillance, predictive patterning and targeted advertising, ushering in a new era in which content delivery was no longer a peripheral feature but the central function of social media’s design.

²⁸ Surabhi Parida, ‘Evolution of Social Media Algorithms: The Invisible Hand Guiding Our Online Experience’ (Reso, 6 October 2024) <https://resoinsights.com/insight/evolution-of-social-media-algorithms/> accessed 15 August 2025.

²⁹ Balkin (n 6) 14.

As platforms transitioned from simple network communities into powerful data-processing hubs,³⁰ the vast volumes of information generated by users provided both the necessity and the raw material for developing advanced content delivery algorithms. Although the specifics of these systems vary across platforms and features, their general logic is simple, involving the use of AI to analyse underlying input signals and predict interactions possibilities.³¹ Relevant signals generally include personal data on users (e.g., their age, location, preferences, past interactions) and data specific to the post being ranked (e.g., its author, the type of media attached, number of likes, author's past performances).³² Relevant interaction possibilities might include the likelihood of a post being liked, commented or shared.³³ Considering these variables, published content is ranked and delivered accordingly.

The introduction of Facebook's "Like" button in 2009³⁴ marks the first milestone in this evolution. By allowing users to express their approval, the "Like" feature enabled Facebook to infer individual preferences and tailor users' Newsfeeds accordingly. Although the initial result helped increase platform use and advance corporate goals, it soon became clear that, not only "Liking," but any measurable interaction—including commenting, clicking, and even hesitating—³⁵ could also be leveraged to advance corporate interests. Regardless of whether they are indicative of a positive or a negative online experience, these interactions reveal precious behavioural data, opening new avenues for platform monetization.³⁶ Realizing this dynamic—and under mounting pressure to demonstrate profitability ahead of its 2012 Initial Public Offering (IPO)—Facebook pioneered systems specifically engineered to predict and provoke user interaction. Not "Liking" specifically, but *any* interaction.³⁷ From then on, the algorithmic

³⁰ Ibid.

³¹ Arvind Narayanan, 'Understanding Social Media Recommendation Algorithms: Towards a Better Informed Debate on the Effects of Social Media' (Knight First Amendment Institute at Columbia, 2023) 5, 6 <http://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms> accessed 27 May 2025.

³² Ibid 18/19.

³³ Ibid.

³⁴ Riemer and Peter (n 8) 412.

³⁵ Meta Newsroom, 'News Feed FYI: Taking Into Account Time Spent on Stories' (Facebook, 12 June 2015) <<https://about.fb.com/news/2015/06/news-feed-fyi-taking-into-account-time-spent-on-stories/>> accessed 11 August 2025.

³⁶ Balkin (n 6) 21; Langvardt (n 3) 19.

³⁷ Balkin (n 6) 21; Riemer and Peter (n 8) 412.

pursuit of engagement superseded users expressed preferences as the primary currency for online visibility.

The term ‘engagement’ generally refers to any measurable user action. Engagement is valuable because it is the simplest and most efficient way through which platforms can directly or indirectly extract information and/or monetary gain to advance their corporate interests. Today, most social media features (Newsfeeds, stories, explore recommendations, etc.)—not just on Facebook, but in all major platforms—are organized “not for chronology, educational value, entertainment value, or expressed user preferences, but for driving ‘engagement.’”³⁸ The meteoric rise of TikTok, with its hyper-personalized, trend-driven “For You” page, epitomizes the latest development of this trajectory.³⁹

Although platforms typically cloak these practices in PR-friendly jargon⁴⁰ that romanticise the supposed benefits of personalization—claiming it exists to “spark conversations and foster meaningful interactions”—⁴¹ their underlying commercial intention is unmistakable. Commercially speaking, optimizing for engagement is a win-win strategy. The longer users scroll through social media, the more likely they are to encounter advertising and eventually made a purchase, thus increasing the platform’s value as an advertising space. Even when users don’t purchase anything, every additional minute of online activity rewards platforms with behavioural data on the type of content that keeps users engaged, fuelling a self-perpetuating cycle of attention extraction.⁴²

More than a commercial strategy, optimizing for engagement has become a cornerstone of the business model operated by social media platforms. So much so that academics have come to define them as “online sites and services that a) host, organize, and circulate users’ shared content or social interactions for them, b) without having

³⁸ Langvardt (n 3) 20.

³⁹ Alex Turvy, ‘Comparing Tiktok and Instagram’s Sociotechnical Environments for Cultural Production.’ (2025) 2 *Platforms & Society* 1 <<https://doi.org/10.1177/29768624251359796>> accessed 20 August 2025.

⁴⁰ Julie E. Cohen, ‘Infrastructuring the Digital Public Sphere’ (2023) 25 *Yale JL & Tech* 1, 17.

⁴¹ Meta Newsroom, ‘Bringing People Closer Together’ (11 January 2018) <<https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/>> accessed 12 of May 2025.

⁴² Cohen (n 40) 17.

produced or commissioned (the bulk of) that content, c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and profit.”⁴³ The aim of this model is not just to learn more about users in order to sell more products. It is also to actively modify user behaviour into generating even more engagement, more data and more revenue opportunities.⁴⁴ To optimize this cycle, platforms have developed structural incentives that simultaneously promote forms of expression that fuel attention extraction and limit those that are inconsistent with their profit-driven imperatives.

As Balkin explains, once media products became important sources of wealth, it was only natural that businesses would seek to concentrate profits through increasingly aggressive forms of legal and technological control.⁴⁵ This resulted in an inevitable conflict: at the very moment when ordinary people are empowered to speak and to participate in the creation of culture, “and to distribute their ideas and innovations around the world, businesses are working as hard as possible to limit and shut down forms of participation and innovation that are inconsistent with their economic interests.”⁴⁶

Although the term ‘moderation’ is usually associated with discrete takedown decisions, in practice, every utterance on social media is ‘moderated.’ Not necessarily to assess its compliance with policy standards, but to determine how far it will circulate, based on its predicted engagement potential. This process affects not only advertisements, but all ‘content’, including news items and personal messages.⁴⁷ If there was ever a time when expression flowed freely, platforms served as mere intermediaries, and virality reflected an idea’s genuine capacity to gain acceptance throughout a connecting network, those days are behind us. The spread of ideas no longer happens organically. The reach and influence of any given publication is determined by AI systems that rank and organize speech according to engagement predictions. These algorithms do not measure the

⁴³ Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018) 18-19 <<https://doi.org/10.12987/9780300235029>> accessed 18 August 2025.

⁴⁴ Riemer and Peter (n 8) 412.

⁴⁵ Balkin (n 6) 14.

⁴⁶ Ibid.

⁴⁷ Riemer and Peter (n 8) 412.

‘value’ of an idea by its truth, coherence, or civic merit, but by its predicted ability to capture user attention.⁴⁸

Chapter 2 –Public Discourse and the Normative Challenges of Algorithmic Content

Distribution

More than influencing individual user behaviour, the shift towards predictive AI in content delivery represents the wholesale commodification of public discourse, where access to audiences becomes contingent on algorithmic viability rather than societal relevance. Platforms are no longer passive intermediaries but active architects of public discourse, with algorithms replacing editorial judgment in determining which ideas register as culturally significant and which fade into obscurity.⁴⁹ The result is a mediated public sphere that naturalizes corporate control over the conditions of democratic deliberation. In his analysis of free speech in the digital age, Balkin captures these dynamics precisely: “[t]he flip side of mass cultural participation is mass personal surveillance, and the danger of widespread digital participation is widespread digital manipulation.”⁵⁰

Platform interference actively influences attention economies, and by extension, political and social agendas. Although social media has famously facilitated significant social movements—such as “#MeToo” in 2017,⁵¹ “Black Lives Matter” in 2020,⁵² and even the Arab Spring uprisings in 2010—⁵³ it has also become associated with the spread of misinformation, polarization, extremist ideologies that spill into violence, and even threats to electoral stability.⁵⁴ This is where the true power of social media lies: in determining what—and who—enters public discourse, and how prominently those messages are amplified.⁵⁵ Through algorithmic systems, platforms shape public

⁴⁸ Ibid.

⁴⁹ Langvardt (n 3) 4.

⁵⁰ Balkin (n 19) 2047.

⁵¹ Fabrizio Botti and others ‘The #Metoo Social Media Effect and its Potentials for Social Change in Europe’ (Foundation for European Progressive Studies, 2019) <https://feps-europe.eu/wp-content/uploads/2019/10/The-Metoo-Social-Media-Effect-and-its-Potentials-for-Change-in-Europe.pdf> accessed 12 August 2025.

⁵² Chang (n 16).

⁵³ Dubai School of Government, *Arab Social Media Report* (Vol 1, No 2, May 2011) <<https://journalistsresource.org/wp-content/uploads/2011/08/DSG-Arab-Social-Media-Report-No-2.pdf>> accessed 15 August 2025.

⁵⁴ See footnotes n 17-22.

⁵⁵ Riemer and Peter (n 8) 410.

discourse at its source, deciding which voices are elevated and which are relegated to the margins.

Algorithmic governance has fundamentally rewritten the rules of public discourse through the silent engineering of visibility itself. The goal of this control is not necessarily censorship, but rather “diversion and co-optation of audience attention.”⁵⁶ Every utterance on social media is filtered through algorithms that determine its reach, influence, and perceived legitimacy. By unilaterally designing these systems, platforms wield unchecked power over which ideas achieve mass visibility and shape public discourse, effectively privatizing the means of mass mobilization. This control creates artificial barriers for participation, translating urgent normative challenges to democracy.

To some extent, contemporary challenges mimic enduring features and concerns that led democracy-based theorists to advocate for traditional mass media regulation in the twentieth century, including:⁵⁷

(1) **Ideological capture and manipulation**, wherein mass media—and now social media—can distort public discourse to advance their own political or economic agendas;

(2) **Epistemic exclusion**, involving the systematic omission of marginalized perspectives, creating knowledge gaps and undermining informed democratic deliberation; and

(3) **Commercial degradation**, where the relentless pursuit of revenue and profit displaces substantive public discourse in favour of sensationalism and mindless entertainment that distracts and demobilizes the public from engaging with pressing civic issues.

However, the digital age compounds these familiar challenges with new concerns that the twentieth-century mindset fails to address. Optimizing for engagement tends to drive users towards strong emotional reactions, including outrage and tribal allegiance, with profound consequences to individuals and societies. As Kayle Langvardt observes:

⁵⁶ Balkin (n 6) 21.

⁵⁷ Ibid 28/29.

“an algorithm designed to drive engagement at all costs is unlikely to push users toward contemplation, intellectual challenge, or doubt. Instead, it will tend to forward users a selection of bias-affirming materials that by turns soothe and provoke the user into more engagement.”⁵⁸ Studies have confirmed this reality, indicating that harmful content, including misinformation and outrage-driven discourse, consistently generates higher user engagement than nuanced or factual material.⁵⁹

By organizing speech distribution around engagement, platforms create a perverse incentive structure where they possess both the technical ability and economic motivation to manipulate the playing field in which ideas compete—often at the expense of deliberation, accuracy, pluralism and user welfare. Rather than fostering deliberation and debate, the attention economy systematically privileges speech that promotes distraction and polarization,⁶⁰ pushing us away from nuanced analysis and consensus-building discourse. This is not a bug but a feature of surveillance capitalism, where platforms function less as Holmes’ idealized public square and more as extractive attention engines.⁶¹

Be it because social media intentionally amplifies harmful and divisive content, as some suggest, or because subjecting speech to profit-driven designs conflicts inherently with human and societal needs,⁶² the consolidation of algorithmic surveillance and predictive targeting in content delivery demands a radical rethinking of free speech protections. When private entities have mastered the power to algorithmically manipulate the marketplace of ideas, the battleground for free expression shifts from a purely formal liberty to speak to preserving the conditions that allow speech to matter in the first place.⁶³

The normative challenges at stake are manifold, but they can be distilled into a central question: how can we counter platform dominance over public discourse and safeguard

⁵⁸ Langvardt (n 3) 21.

⁵⁹ Vosoughi, Roy and Aral (n 20); Laura Edelson and others, ‘Far-Right News Sources on Facebook More Engaging.’ (Medium, 2021) <<https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging04a01efae90>> accessed 21 February 2025.

⁶⁰ Langvardt (n 3) 2.

⁶¹ Cohen (n 9) 70.

⁶² Narayanan (n 31) 35.

⁶³ Balkin (n 6) 48/49.

free speech against the distortive commodification pressures embedded in social media's content distribution systems?

Chapter 3 – A Brief Review of Literature on 'New School' Speech Regulation and its Gaps

When explaining the evolution of speech governance in the algorithmic society, Balkin distinguishes between 'old school' and 'new school' speech regulation. The former, he explains, targets individual speakers and media companies through censorship, prior restraint, or criminal penalties. New school regulation, however, is directed at internet intermediaries and aims at coercing and co-opting digital infrastructures into policing and controlling online speech.⁶⁴ New school regulation depends on private governance; meaning it depends on the ability of infrastructure providers to govern the flow of information and the behaviour of end-users who employ their digital infrastructure.⁶⁵ This ability derives from telecommunications law, property rights, and from contractual agreements between infrastructure providers and end-users.⁶⁶

The methods developed by twenty-first century technologies to control online speech are vast. They include blocking, filtering, removing content, controlling visibility, controlling access to digital facilities, up ranking or downranking, shadow banning, flagging, monetization or demonetization and general algorithmic surveillance.⁶⁷ However, whenever private governance is channelled through regulation—or even when it is discussed more broadly—debates tend to focus on a perceived trade-off between online safety and private censorship. The EU's E-Commerce Directive demonstrates the narrowness of early new school frameworks by addressing platform control entirely through the lens of content takedowns, instituting procedural requirements like transparency and due process for removal.⁶⁸ Meanwhile, the far more

⁶⁴ Balkin (n 19) 2015.

⁶⁵ Jack M. Balkin 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 *UC Davis Law Review* 1148, 1182 <<https://openyls.law.yale.edu/server/api/core/bitstreams/2a57c714-e278-449a-8ba0-91c6d56c34e2/content>> accessed 20 August 2025.

⁶⁶ *Ibid.*

⁶⁷ *Ibid* 1175; Cohen (n 40) 10.

⁶⁸ Tambiama Madiaga, *Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act* (European Parliamentary Research Service, May 2020) PE 649.404 <[https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA\(2020\)649404_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA(2020)649404_EN.pdf)> accessed 15 June 2024.

pervasive and consequential mechanisms employed to algorithmically determine the visibility of speech—regardless of perceived harmfulness—were left unaddressed.⁶⁹

In 2019, Tim Wu described the invisible dimension of algorithmic speech control in a paper that examines the future of social ordering systems. In it, he distinguishes between ‘positive’ (or ‘affirmative’) and ‘negative’ tools of speech control. Positive control “entails choosing what is brought to the attention of the user”—through algorithmic curation, recommendations, and the promotion or demotion of search results.⁷⁰ Conversely, negative control involves the reactive removal of rule-violating content. Later, Enrique Armijo and other scholars proposed reclassifying this distinction as ‘Type I’ moderation, referring to the algorithmic ordering and ranking of content display, and ‘Type II’ moderation, referring to the detection, assessment, and removal of offending content. The terminological update was defended on the grounds that ‘positive’ moderation discriminated content both positively and negatively⁷¹—simultaneously amplifying to drive engagement, and suppressing content deemed low quality or less commercially viable.

While these works approach an understanding of the structural dimension of speech control on social media beyond the well-established problematics of content removal, they fall short of exploring its social and normative implications. Wu concludes his taxonomy by asserting that content removal remains “inherently more controversial”⁷² than the structural mechanisms operated to deliver content, choosing to focus on the former rather than the latter. Similarly, Armijo also chooses to prioritize content removal, arguing that the manipulative potential of algorithmic content distribution would render it theoretically “more amenable to regulation to prevent or minimize harms associated with its use”⁷³ when compared to its counterpart.

⁶⁹ Riemer and Peter (n 8) 419.

⁶⁹ Ibid 412.

⁷⁰ Tim Wu, ‘Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-ordering Systems’ (2019) 119 *Columbia Law Review* 2001, 2014. <<https://www.jstor.org/stable/10.2307/26810857>> accessed 15 March 2025.

⁷¹ Enrique Armijo, ‘Speech Regulation by Algorithm’ (2021) 30 *William & Mary Bill of Rights Journal* 245, 248. <<https://scholarship.law.wm.edu/wmborj/vol30/iss2/3>> accessed 15 March 2025.

⁷² Wu (n 70) 2014.

⁷³ Armijo (n 71) 249.

By framing content moderation predominantly as a problem of removal, these cases exemplify a wider scholarly tendency that creates a polarized yet narrow debate on content moderation.⁷⁴ On one side are those who criticize platform inaction and demand stronger efforts against harmful content; on the other are those who warn against corporate overreach and call for stricter oversight of removals. Comparatively, algorithmic content distribution is featured as a less urgent, or less attainable, problem, when it is addressed at all. This issue is acknowledged by Julie E. Cohen in her assessment of the various processes and priorities that constitute content governance on social media. The lack of readily available frames of reference in recursive free speech thinking, she argues, has prevented more sophisticated discussions on content governance. As a result, pressing issues have been sidelined in favour of arguments that focus narrowly on the harms of censorship in content moderation.⁷⁵

Keeping public and policy debate fixated on matters of censorship has worked to the advantage of social media platforms, allowing them to sidestep serious scrutiny of the monetization logics that underline their business models.⁷⁶ This framing has been reinforced by platforms themselves, who work hard “to ensure that the nature of the predictive patterning they impose does not become (or remain) a focus for the public and for policymakers.”⁷⁷ An example of this is Facebook’s public plea for stricter regulation against ‘harmful content’ following its highly mediatic 2019 congressional hearing in the aftermath of the Cambridge Analytica scandal.⁷⁸ By expressing its supposed concern and asking public authorities to take a stronger stance to help control the spread of harmful content, Facebook subtly steers attention away from the fact that visibility and virality are direct products of its own engagement-driven policies and algorithms.

This reality reveals a clear and present danger: as long as debates remain framed in current terms, the structural mechanisms that determine which expressions are amplified and which are marginalized will continue to evade scrutiny. Such evasion

⁷⁴ Riemer and Peter (n 8) 419.

⁷⁵ Cohen (n 40) 18.

⁷⁶ Riemer and Peter (n 8) 420.

⁷⁷ Ibid.

⁷⁸ ‘Mark Zuckerberg asks governments to help control internet content’ (*BBC News*, 30 March 2019) <<https://www.bbc.com/news/world-us-canada-47762091>> accessed 14 July 2025.

implicitly legitimizes these practices,⁷⁹ entrenching them and leaving little scope for challenge in the future.

Recognizing this threat, a small but influential group of scholars has managed to advance discussions on the normative and regulatory challenges of algorithmic content distribution. In 2021, Riemer and Peter published a provocative pre-theory investigation centred on the structural influence of algorithms in audience access, a phenomenon they term 'algorithmic audiecing.'⁸⁰ Their inquiry identifies audience determination as a core mechanism of user and discourse control and exposes the monetization logic that determines whose speech is amplified or suppressed as a key focus point for future digital regulation.⁸¹ Moreover, their investigation correctly acknowledges that, because they affect "all speech, routinely and always," accounting for the mechanisms that control visibility on social media is more important than expanding current content removal regimes. Whereas content removal "pertains only to speech that is perceived as harmful or otherwise problematic, thus a small proportion of all messages, all speech is subjected to algorithmic content distribution."⁸²

Cohen's diagnosis that the worst pathologies of social media stem from the "synergies between pattern-driven, predictive targeting and processes of social circulation platforms,"⁸³ resonates with Riemer and Peter's conclusions, underscoring the urgency of disciplining these dynamics. Coupled with her extensive analyses of the operational logic of surveillance capitalism, which illuminates the causes and consequences of algorithmic content distribution, her work has proven foundational to this study.

Significant contributions also emerge from Kayle Langvardt's scholarship. His interpretation of social media platforms as 'speech regulators', underscoring how the use of AI has allowed "a new kind of power over cultural flows,"⁸⁴ generating risks of manipulation and distortion, is particularly insightful. Like Riemer and Peter, Langvardt

⁷⁹ Riemer and Peter (n 8) 422.

⁸⁰ Ibid.

⁸¹ Ibid 410.

⁸² Ibid 419.

⁸³ Cohen (n 40) 38.

⁸⁴ Langvardt (n 3) 4.

argues that these dynamics must be addressed through regulation that intentionally targets platforms' business models, proposing concrete measures to achieve this.

Taken together, these contributions lay a crucial foundation for reframing the debate on new school speech regulation beyond narrow concerns of censorship and liability. A more comprehensive approach must account for the systemic ways in which platforms shape expression—not just through takedowns—but structurally and proactively, through content distribution policies and algorithmic design. By interrogating the economic drivers of speech distribution and emphasizing the need for interventions that target them directly, the reviewed literature marks an important advance

Yet, despite strides, scholarship remains fragmented and limited. Riemer and Peter's investigation focuses narrowly on the issue of audience access, overlooking the broader normative issue of subjecting speech and public discourse to profit-driven metrics. Although they call for structural interventions, they neither explain why existing frameworks are inadequate nor propose concrete regulatory measures, leaving a gap between critique and actionable reform. Langvardt, by contrast, offers compelling suggestions for specific regulatory action. However, these proposals are not anchored in a systematic theory of free speech making it difficult to measure normative coherence and regulatory success. Without this theoretical grounding, his interventions risk appearing *ad hoc* rather than part of a unified vision for protecting free expression in the digital age. Finally, Cohen's analysis approaches platform dysfunctions broadly, lacking details that are pertinent to the protection of speech and public discourse specifically.

Missing from the academic landscape is a theory of 'new school' regulation that not only recognizes the structural dimensions of platform interference but is also practically equipped to confront its profound democratic deficits, all while being grounded in a coherent vision of free speech protection and democratic participation.

Part II - A Critique of EU Regulation

Preliminary Considerations: A Brief Justification

In the previous section, we have outlined a missing theoretical framework for algorithmic governance and free speech—one that addresses the structural dimension

of platform control and grounds structural solutions in democratic and free speech values. This gap is particularly pressing considering the current trajectory of EU regulation.

In recent years, growing concerns over the power asymmetries of AI deployment and platform governance have triggered overdue regulatory response. This culminated in the adoption of landmark legislation in the EU: the Digital Services Act (DSA) and the Digital Markets Act (DMA) in 2022, followed by the Artificial Intelligence Act (AI Act) in 2024. Owing to the so-called Brussels Effect, these frameworks are frequently viewed as regulatory blueprints on matters of AI and platform governance, influencing policies well beyond the EU. Despite this, while they occasionally acknowledge the threats of algorithmic interference in free speech and public discourse, these frameworks fail to provide an adequate response to the normative challenge of countering the distortive effects of content distribution systems.

The AI Act's surprising omission of the algorithms that govern content distribution, the critical disjuncture between the structural nature of the problem and the procedural focus of the DSA, and the missed opportunity to address the broader societal impacts of platform control over ranking systems in the DMA—ultimately leave the question unresolved.

Yet, it remains important to understand how and why these frameworks fail, for it is precisely in these failures that a path forward emerges, revealing the contours of a more direct and substantial solution—one that must confront the structural incentives driving algorithmic governance, redefine platform accountability, and reclaim democratic control over online discourse.

The following chapters will address the individual and systemic vulnerabilities of the EU's current regulatory paradigm on AI and platform governance, highlighting the loopholes, blind spots, and methodological flaws that hinder it ineffective against the economic forces corrupting online discourse. This analysis does not intend to provide comprehensive and definitive interpretations of these very recent frameworks. Rather, it seeks to expose the gaps that require bridging and the regulatory misconceptions that must be revised to ensure meaningful protection against the structural dimensions of

platform control. Through this critical examination, we aim to distil actionable insights for future regulatory interventions—ensuring they are deliberately designed to realign AI deployment and platform governance with democratic interests. Without such a holistic understanding, reforms risk continuing to treat symptoms rather than root causes, leaving platform power unchecked and democratic discourse vulnerable to corporate interests.

Chapter 1 - The AI Act: An Attempt to Interpret Potential Impacts Amidst Legislative Silence

Enacted on 21 May 2024, the EU's AI Act establishes a risk-based framework that classifies AI systems into four tiers—unacceptable, high, limited, and minimal risk—each subject to a distinct set of obligations. Social media platforms rely heavily on AI to train and refine the systems that automate content distribution on users' feeds.⁸⁵ Yet—despite extensive public debates over their negative effects on user welfare and societal stability—these systems are conspicuously absent from the Act's explicit categories.

Although this omission has attracted criticism from specialized scholars,⁸⁶ its repercussions remain limited, calling for broader and more sustained engagement within the legal community. One of the most pressing issues open for debate is whether this silence reflects a legislative oversight—open to corrective interpretation by courts and regulators—or an intentional political choice to exempt algorithmic content distribution from stringent oversight. The implications of this uncertainty, and its possible causes, will be further addressed in future chapters (see considerations on the bottleneck effect of EU regulation in Chapter 3). For now, we will attempt to interpret what implications the AI Act might have on content distribution, amidst its legislative silence.

⁸⁵ As acknowledged by Meta: Meta Transparency Centre, 'Explaining Ranking on Facebook' (Transparency Centre, 31 December 2023) <https://transparency.meta.com/features/explaining-ranking/#facebook_ai_systems> (31 of December 2023) accessed 16 of May 2025.

⁸⁶ Particularly: Sandra Wachter, 'The Limitations of the EU's Artificial Intelligence Act and the AI Liability Directive' (2024) 26 *Yale Journal of Law & Technology* 674, 711 <https://yjolt.org/sites/default/files/wachter_26yalejltech671.pdf> accessed 14 May 2025.

A difficult case for prohibition under article 5(1)

Article 5(1) of the AI Act establishes a cross-sectorial ban on practices considered to pose unacceptable risk. Although the ranking, sorting, and distribution of content are not explicitly mentioned, a case could be made for significant restrictions of current industry practices under Article 5(1)(a)—which prohibits systems from employing “purposefully manipulative techniques to distort user behaviour.” Constraints could also be argued under Article 5(1)(b)—which bans the exploitation of vulnerable groups—and Article 5(1)(c)—which prohibits the use of AI for systematic scoring of users based on personal data, personality traits or predicted behaviour.

Such interpretations are plausible given that platforms have publicly acknowledged practicing behavioural scoring.⁸⁷ Furthermore, an increasing body of research has linked AI-driven architectures to psychological manipulation, behavioural addiction, and measurable harm in children, adolescents, and groups with mental health vulnerabilities.⁸⁸

However, enforcing reform to standard industry practices under the Act’s prohibition regime is not without its challenges. Notably, Recital 29 explicitly exempts “common and legitimate commercial practices, for example in the field of advertising” from prohibition under Article 5(1). Given that platforms have employed engagement-driven AI to rank, select and deliver speech for over a decade, and that this practice has been frequently associated with advertising (including by platform themselves), this carve-out—

⁸⁷ Particularly, the infamous 2014 emotion study where Facebook subjected 689,003 users to manipulated news feeds of users to see if they could influence their emotions, demonstrating that user behavior could be tracked and altered. The case was widely reported in mainstream media, including: Kashmir Hill, ‘Facebook Manipulated 689,003 Users’ Emotions For Science’ (*Forbes*, 28 June 2014) <<https://www.forbes.com/sites/kashmirhill/2014/06/28/facebook-manipulated-689003-users-emotions-for-science/>> accessed 14 July 2025; Vinu Goel, ‘Facebook Tinkers With Users’ Emotions in News Feed Experiment, Stirring Outcry’ (*The New York Times*, 29 June 2014) <<https://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html>> accessed 14 June 2025.

⁸⁸ On the harms of social media to children and other vulnerable groups: *Pathways: How Digital Design Puts Children at Risk* (5Rights Foundation, July 2021) <<https://5rightsfoundation.com/wp-content/uploads/2021/09/Pathways-how-digital-design-puts-children-at-risk.pdf>> accessed 14 August 2025; Maya Brownstein, ‘Social Media Platforms Generate Billions in Annual Ad Revenue from U.S. Youth’ (Harvard T.H. Chan School of Public Health, 27 December 2023) <<https://hsph.harvard.edu/news/social-media-platforms-generate-billions-in-annual-ad-revenue-from-u-s-youth/>> accessed 14 August 2025; *Protecting Youth Mental Health: The U.S. Surgeon General’s Advisory* (US Department of Health and Human Services, 2021) <<https://www.hhs.gov/sites/default/files/surgeon-general-youth-mental-health-advisory.pdf>> accessed 14 August 2025.

predicated on the nebulous standard of ‘legitimate commercial practice’—creates a loophole that platforms will likely exploit for years before judicial interpretation catches up.⁸⁹

By failing to regulate the industry’s exploitative engagement-driven practices under Article 5(1) explicitly, the AI Act missed a critical opportunity to curb harmful design tactics and recalibrate them with human and societal needs—undermining the regulation’s promise of promoting “human-centric and trustworthy AI.”

Between high-risk and systemically risky GPAI: uncertain roads to platform accountability

Given the challenges of regulating algorithmic content distribution under Article 5(1)—and considering that systems deemed to pose minimal risk face only superficial transparency requirements that would hardly apply to content distribution (e.g., disclosing AI-generated content)—⁹⁰ the AI Act leaves only two possible paths for platform accountability. The first is Article 6’s high-risk regime, the second Article 51’s framework for systemically risky general-purpose AI (GPAI).

Under the high-risk classification, systems are subjected to seemingly strict requirements, including risk management (Article 9), record-keeping (Article 12), mandatory human oversight (Article 14), and post-market monitoring (Article 61). Unlike Article 5’s cross-sectorial approach to defining unacceptable practices, Article 6 adopts a rigid sector-specific method of classifying systems as high-risk with affected areas being listed under Annex III.⁹¹ This list includes healthcare, immigration and law enforcement. Once again, despite extensive social debates and well-documented potential risks, the use of AI to rank, select and deliver speech on social media is not mentioned.⁹² Even worse, Article 6’s closed-list methodology prevents interpreters from

⁸⁹ On the harms of digital market manipulation: Ryan Calo, ‘Digital Market Manipulation’ (2014) 82 *George Washington Law Review* 995, 1026 <<https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=1024&context=faculty-articles>> accessed 14 August 2025.

⁹⁰ Wachter (n 86) 684.

⁹¹ Martin Ebers, ‘Truly Risk-based Regulation of Artificial Intelligence How to Implement the EU’s AI Act’ (2024) 15 *European Journal of Risk Regulation* 1, 9 <<https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/truly-riskbased-regulation-of-artificial-intelligence-how-to-implement-the-eus-ai-act/E526C1D0D7368F9691082220609D60F4>> accessed 14 May 2025.

⁹² Wachter (n 86) 716.

conducting the necessary case-by-case analysis⁹³ that might otherwise recognize these systems as high-risk given their demonstrable impacts on fundamental rights.

Under Article 7(1), the European Commission has the authority to amend Annex III to include new applications within the high-risk regime. To do so, it must verify both harm to fundamental rights and the inadequacy of existing risk-mitigation measures. Given the well-documented social repercussions of content-ranking systems—from the suppression of marginalized voices to the amplification of hate speech and the broader erosion of democratic discourse—this provision offers some hope that such systems may eventually be brought within the Act’s high-risk regime. Yet, this analysis raises a critical question: If the amendment process requires proof of harm and regulatory failure, were these very factors not considered during the Act’s initial drafting?⁹⁴ If so, what could possibly justify the exclusion of such controversial and pervasive systems not only from the high-risk classification, but from the scope of the AI Act altogether? These and other relevant provocations will be further analysed in Chapter 4.

Finally, Article 55 of the AI Act regulates systemically risky GPAI, subjecting providers to assessments, mitigation strategies, record-keeping, incident reporting and ensuring adequate levels of cybersecurity. The ‘systemic risk’ threshold is defined in Recital 110 as encompassing severe harm to public health, public safety, democratic processes, or “the spread of harmful content.” This final reference offers a potentially promising pathway for regulating content distribution under the systemically risky GPAI regime. However, ultimately, this promise is undercut by the lack of legal clarity regarding whether these systems qualify as GPAI in the first place. This uncertainty arises from the Act’s failure to provide a binding definition of GPAI—an oversight that the Commission itself explicitly acknowledged in 2025.⁹⁵

The absence of a legally binding definition effectively shifts a fundamentally legislative issue into the hands of regulators and courts, likely resulting in years of legal uncertainty and inconsistent application. While proponents may argue that content distribution

⁹³ Ebers (n 91) 10.

⁹⁴ Ibid 9.

⁹⁵ European Commission, ‘Commission Seeks Input to Clarify Rules for General-Purpose AI Models’ (Press Release, 22 April 2024) <<https://digital-strategy.ec.europa.eu/en/news/commission-seeks-input-clarify-rules-general-purpose-ai-models>> accessed 14 May 2025.

systems serve multiple functions—including surveillance, ad targeting and content moderation—platform enthusiasts are likely to counter that these systems are platform-specific and are generally designed for the sole purpose of optimizing engagement. This ambiguity, compounded by the lack of scholarly consensus on what constitutes GPAI, leaves the path for platform accountability uncertain. As a result, content distribution systems may end up facing only superficial transparency obligations—such as disclosing AI-generated content—required of minimum risk AI under Article 50.⁹⁶

Such a minimalistic regime fails to advance the protection of free speech and public discourse against systematic distortions in any meaningful way. Its sole aim is distinguishing human-made from machine-generated content, remaining entirely agnostic to the nature, spread, and societal impact of such content. Consequently, it proves inadequate in addressing the normative challenge of countering platform dominance over public discourse and safeguarding free speech against the distortive effects of their business model.

Chapter 2 - The DSA's Attempt to Regulate Structural Controls With Superficial Procedural Rules

A landmark for addressing content distribution systems

Enacted in 2022, the Digital Services Act (DSA) materializes the EU's efforts to recalibrate power asymmetries between platforms and online service users, imposing obligations mainly related to content moderation, algorithmic transparency, and user empowerment. It builds upon the E-Commerce Directive, modernizing liability frameworks while introducing stricter requirements for Very Large Online Platforms (VLOPs).⁹⁷ Unlike the AI Act, the DSA addresses content distribution systems specifically, representing a landmark in EU legislative history.

As previously discussed, the E-Commerce Directive addressed only reactive mechanisms of content control (takedowns). Recognizing the existence of a deeper, structural architecture of proactive control—one that holds a significant role in the “amplification

⁹⁶ Wachter (n 86) 684.

⁹⁷ Miriam C. Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation' (2021) 12 *JIPITEC* 361, 371 <<https://www.ssrn.com/abstract=3876328>> accessed 14 June 2025.

of certain messages” and “in stimulating online behaviour” (Recital 70)—the DSA inaugurates a regime to target these so-called ‘recommender systems.’

Article 3(s) defines these systems as fully or partially automated mechanisms that determine the visibility, ranking, or display order of content presented to users. Throughout its framework, the DSA requires VLOPs to disclose the main parameters governing recommendation algorithms (Article 27), provide non-profiling-based alternatives (Article 38), and conduct systemic risk assessments that account for the role of these systems in disseminating harmful content (Articles 34-35). Additionally, VLOPs are required, under Article 35, to adjust the design of their recommender systems where necessary to prevent and minimise systemic risks.

While these provisions directly address the structural controls employed for content distribution, their ability to meaningfully advance the protection of free speech and democratic discourse from distortion is questionable. Rather than enhancing observability and accountability over platform practices, the measures the DSA employs enable platforms to selectively curate disclosures that technically meet its requirements while continuing to withhold data capable of informing substantive policy interventions.⁹⁸

This calculated opacity is not an oversight, but a hallmark of digital regulation, perpetuating a regulatory approach that allows platforms to perform compliance while keeping underlying harmful practices unchanged. The result is a regulatory regime that prioritizes procedural box-checking over substantive oversight and accountability—an approach that is fundamentally unfit to advance the protection of democratic participatory values and fundamental free speech rights.

The limits of transparency: why disclosure regimes are unreliable, ineffective and fail to govern dynamic algorithmic systems

In their work on algorithmic transparency and platform governance, Anna-Katharina Meßmer and Martin Degeling argue that the aim of transparency and data access

⁹⁸ Regarding the insufficiency of transparency oriented regimes for AI governance, see: Woodrow Hartzog, ‘Two AI Truths and a Lie’ (2023) 26 *Yale Journal of Law & Technology* 596, 631 <https://vjolt.org/sites/default/files/hartzog_26yalejltech595.pdf> accessed 14 August 2025; Narayanan (n 31) 22.

obligations should be to “maximise observability”—defined as our sustained capacity to scrutinise platforms over time in order to “detect their evolution and impacts on individuals, groups and society (...).”⁹⁹ The goal of these mandates is, therefore, not to reform industry practices necessarily but to create official records that might help society understand their long-term effects. Yet, a closer examination of how speech-regulating infrastructures work, and what DSA-mandates disclosures have effectively delivered into public domain, reveals that even this very humble objective has not been achieved.

As of July 2025, Meta remains the only major company providing actively updated, detailed documentation on its ranking and recommendation algorithms. By contrast, X¹⁰⁰ and TikTok’s¹⁰¹ DSA-mandated reports primarily aggregate raw data on the removal of illegal content, offering little systematic insight into their systems’ architecture and decision-making processes. For this reason, our analysis will focus on Meta’s systems and the information made available in its DSA-mandated report and its Transparency Centre.¹⁰²

In detailing its systems, Meta acknowledges using “over 100 interdependent prediction models” to analyse “thousands of different signals”—ranging from user behaviour to content context.¹⁰³ The influence of each model on an individual’s feed is dynamic and unpredictable, varying according to personal preferences, history and interaction patterns. As Meta states, these models are “continuously updated to deliver the most relevant experience”¹⁰⁴ to users. While their exact configurations are constantly changing, frequently used models may include:

⁹⁹ Anna-Katharina Meßmer and Martin Degeling, ‘Auditing Recommender Systems -- Putting the DSA into practice with a risk-scenario-based approach’ (Stiftung Neue Verantwortung, February 2023) 2, 55 <<https://doi.org/10.48550/arXiv.2302.04556>> accessed 14 May 2025.

¹⁰⁰ X Corp., *Global Transparency Report: January - June 2024* (February 2025) <<https://transparency.x.com/en/reports/global-reports/2025-transparency-report>> accessed 14 May 2025.

¹⁰¹ TikTok Inc., *Community Guidelines Enforcement Report: First Half 2025* (26 June 2025) <<https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2025-1>> accessed 14 of May 2025.

¹⁰² Meta Transparency Centre, ‘Our Approach to Facebook Feed Ranking’ (Meta Transparency Centre, updated 11 June 2025) <<https://transparency.meta.com/pt-br/features/ranking-and-content/>> accessed 27 August 2025.

¹⁰³ Ibid.

¹⁰⁴ Ibid.

- How likely you are to be interested in content from your friends
- How likely you are to be interested in the Group that shared the post or content from the Group, as measured by engagement with the Group or its content
- How likely you are to be interested in the Page that shared the post or content from the Page, as measured by engagement with the Page or its content
- How likely you are to click on some part of the post
- How likely you are to interact with a post in some way by liking, reacting or commenting on it
- How likely you are to meaningfully interact with the post, through some combination of commenting/liking/reacting/sharing to messenger/resharing and/or spending time viewing it
- How likely you are to share the post
- How likely you are to spend time viewing comments on the post
- How likely you are to spend time viewing the post or content in the post (as opposed to just scrolling past it)
- How likely you are to visit a Page after seeing a post from that Page
- How likely you are to want to see more or see less content from the person or Page who shared the post
- How likely you are to watch a video contained in the post and the predicted amount of time you'll spend watching it
- How likely your interactions on the post from a Page will encourage the admin of that Page to share more content in the future that is valuable to you
- The predicted amount of additional comments or replies a post will get if you comment or share the post

While Meta's disclosures may technically comply with the DSA vague “main parameters” requirement, they fail to meaningfully enhance the observability of platforms’ content delivery practices, frustrating the standard previously proposed to evaluate

transparency mandates. Three fundamental flaws explain this failure—one rooted in the DSA's design, another inherent to transparency regimes generally, and a third specific to the algorithmic systems that govern content distribution.

First, the DSA grants VLOPs excessive discretion in determining what might constitute 'sufficient' transparency. This excessive leeway allows Meta and other companies to cherry-pick the aspects they want to make public and claim trade secrecy over the rest.¹⁰⁵ Unsurprisingly, Meta frames Facebook's and Instagram's recommendation systems in PR-friendly jargon,¹⁰⁶ romanticizing the benefits of 'personalization' while outright obscuring its harmful effects. As renowned computational scientist and Princeton University professor Arvind Narayanan notes, "companies have shared precious little about the *effects* of algorithms."¹⁰⁷ While this is precisely the kind of data needed to inform evidence-based policies, platforms still seem to have little incentive to disclose it—even after the DSA.¹⁰⁸

Second, the fundamental disconnects between corporate rhetoric, past misconduct and generally observable outcomes expose the inherent limitations of transparency mandates as effective regulatory tools. While platforms often downplay the negative effects of their governance models—claiming ignorance, or that they are doing their best to mitigate harms—investigative reporting has frequently revealed incompatibilities between these claims and internal practices.¹⁰⁹

The 'Facebook Files' series, an investigative piece published by *The Wall Street Journal* in 2021, exemplifies this contradiction. Based on leaked internal documents, the series uncovers misconducts that range from deliberately shielding high-profile users from

¹⁰⁵ Narayanan (n 31) 36; Cohen (n 40) 10.

¹⁰⁶ For example: "We make personalised recommendations to the people who use our services to help them discover new communities and content. (...) Our goal is to make recommendations that are relevant and valuable to each person who sees them. We work towards our goal by personalising recommendations, which means making unique recommendations for each person." In: Facebook Help Centre, 'How does Facebook rank posts in Feed?' Facebook Help Centre (updated 11 June 2024) <<https://www.facebook.com/help/1257205004624246/>> accessed 14 August 2025.

¹⁰⁷ Narayanan (n 31) 22.

¹⁰⁸ Cohen describes this unwillingness in her assessment of platform disclosures: "External observers are forced to rely on those disclosures—which are mostly about results and to some extent also about reasons—because no intermediary has provided operational detail about its algorithmic moderation processes, and none has been willing to allow interrogation of underlying practices of optimizing for immoderation in the first place." In: Cohen (n 9) 250.

¹⁰⁹ Cohen (n 40) 32.

community guidelines (even when their content incited harassment or violence)¹¹⁰, to omitting knowledge on Instagram’s harmful effects on teenage girls before the United States Congress.¹¹¹ The series has also exposed CEO Mark Zuckerberg’s resistance to reversing changes that made the platform, and its users, measurably “angrier” in 2018 because doing so would likely reduce engagement.¹¹²

This example exposes the inherent limitation of transparency mandates. No matter how strict these obligations get, companies will always have an overwhelming incentive to omit critical information, curate favourable data narratives, and exploit legal loopholes to leave their business models intact.¹¹³ This fundamental misalignment of corporate and public interest goals ensures that, as long as regulations continue to focus on transparency, compliance will remain performative and accountability illusory.¹¹⁴

Finally, transparency mechanisms are inherently ill-suited to govern complex and dynamic systems like the ones that organize users’ feeds. As informed by Meta, these systems synthesize “hundreds” of interdependent predictions, derived from “thousands” of input signals. The sheer amount of data is impossible to imagine, let alone subject to reliable oversight—even if systems were not continuously evolving and refining (which they are).¹¹⁵ The decision-making processes of these systems are so complex and so dynamic that their specific pathways become fundamentally non-replicable and opaque—even to their operators. Opacity is not incidental, but a fundamental feature of algorithmic design which will not be minimised by regulation that continues to insist on transparency, regardless of how strict these obligations get.¹¹⁶

¹¹⁰ Jeff Horwitz, ‘Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That’s Exempt.’ *The Wall Street Journal* (13 September 2021) <https://www.wsj.com/tech/facebook-files-xcheck-zuckerberg-elite-rules-11631541353?mod=article_inline> accessed 14 August 2025.

¹¹¹ Georgia Wells and others, ‘Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show.’ *The Wall Street Journal* (13 September 2021) <https://www.wsj.com/tech/personal-tech/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?mod=hp_lead_pos7,article_inline> accessed 14 August 2025.

¹¹² Keach Hagey and Jeff Horwitz ‘Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead.’ *The Wall Street Journal* (15 of September 2021) <https://www.wsj.com/tech/facebook-algorithm-change-zuckerberg-11631654215?mod=article_inline> accessed 14 August 2025.

¹¹³ Hartzog (n 98) 631.

¹¹⁴ Cohen (n 40) 12.

¹¹⁵ Cohen (n 9) 247.

¹¹⁶ Narayanan (n 31) 35.

The fox guarding the henhouse: the failures of industry self-assessment

The DSA's risk assessment mandate (Articles 34-35) could, in principle, provide a more robust mechanism for platform accountability compared to transparency requirements alone. Yet, its transformative potential is critically undermined by a series of structural deficiencies that entrench platform discretion and reinforce industry incentives to prioritize performative compliance over genuine evaluation and reform. First, assessments are conducted unilaterally and are not subject to external oversight or validation. Second, the DSA fails to establish standardized benchmarks for risk assessment or thresholds for risk mitigation. Finally, it does not specify methodological requirements, giving platforms' full discretion on how to conduct their own evaluations.

Meta's 2023 Risk Assessment Report for Facebook,¹¹⁷ illustrates these issues. It covers the four categories of systemic risks outlined in Recitals 80-83 of the DSA, including: (1) risks associated with the dissemination of illegal content; (2) actual or foreseeable impacts on the exercise of fundamental rights, including freedom of expression and of information; (3) actual or foreseeable negative effects on democratic processes, civic discourse and public security; and (4) actual or foreseeable negative effects on public health or on gender-based violence.

While the report acknowledges the damaging potential of Facebook's systems to "fundamental rights" and "civic discourse" specifically (pages 19 and 26 of the report), analysis remain superficial, with systemic harms being framed as either misuse or inevitable trade-offs rather than addressable issues. A striking example of this imbalance is how the assessment focuses primarily on content-related risks, framing risk around the recommendation of potentially undetected problematic content, while overlooking the contribution of Facebook's own engagement-driven policies and designs (page 29 of the report).

The methodology used in the actual risk assessment is also questionable. It employs an opaque 'Tier' classification to boast the effectiveness of existing controls, while providing no empirical evidence to validate these results. For example, Meta claims to

¹¹⁷ Meta, *Systemic Risk Assessment and Mitigation Report for Facebook* (Pursuant to Regulation (EU) 2022/2065, August 2024) <<https://transparency.meta.com/reports/regulatory-transparency-reports/>> accessed 20 May 2025.

have reduced ‘Residual Risk’ across all ‘Problem Areas’—including misinformation, disinformation, and hate speech—to Tier 1 and Tier 2 levels from 2022 to 2023 (page 42 of the report). This would mean that the current risk levels of misinformation, disinformation and hate speech on Facebook are either ‘extremely low’ or ‘low.’ Not only does this claim seem exaggerated, but it also lacks methodological accuracy and empirical validation.

Meta’s Residual Risk calculation depends on two elements: ‘Inherent Risk’, representing the baseline risk without Facebook’s mitigation measures, and ‘Control Effectiveness Reduction’ (CSE%), representing the claimed impact of its mitigation measures. The Inherent Risk score is approached through two scales, a ‘Severity’ rubric (which considers estimated scale of impact¹¹⁸, nature of the impact¹¹⁹, and impacted demographic)¹²⁰ and a ‘Likelihood’ rubric (evaluating scope¹²¹ and volume¹²² of impacted users). Meanwhile, CSE% is asserted through subjective and qualitative judgments, comprising criteria like whether safeguards are “robustly designed” or “consistently applied”, as illustrated on page 92 of the report:

Purpose	Is the control detective or preventive, and is it automated or manual?
Appropriateness	Is the control built in a robust manner to mitigate the risk?
Frequency & Consistency	Does the control occur with appropriate frequency and is applied in a consistent manner?
Aggregation	Is the control designed at the correct level of the organisation to mitigate the risk?
Investigation	Can the control be evaluated and investigated in the future? Is there appropriate documentation?
Dependency	Does the control have a reliant design and built in a sustainable manner?
Competency	Does the operator of the control have the right skill level to operate the control?

The ‘Residual Risk’ output is obtained by multiplying ‘Inherent Risk’ by the CSE%, as illustrated below (page 90 of the report):

¹¹⁸ Ranging from, Tier 1: “isolated to Meta systems” to, tier 5: “severe ramifications to healthcare or financial sector, broader society stability and protection of individuals, including marginalized groups.”

¹¹⁹ Ranging from, Tier 1: “general Meta user experience” to, Tier 5: “impact to fundamental rights, including freedom of expression.”

¹²⁰ Ranging from, Tier 1: “repeat offenders” to Tiers 4 and 5: “society at large” and “minors and marginalized groups” respectively.

¹²¹ Ranging from, Tier 1: 250k used to, tier 5: >200M users.

¹²² Ranging from, Tier 1: “risk is rarely seen by users and must be deliberately searched” to, Tier 5: “risk is persistently seen and is found inadvertently.”



While Meta's dual-axis, multi rubric, approach appears systematic, the lack of transparency regarding specific scoring mechanisms, underlying data, and clear justifications for how results correlate with observable outcomes prevents independent verification of the company's findings. Without access to the exact rubrics, weightings, and evidence supporting each assessment, Meta's self-assessed methodology is fundamentally prone to bias and misrepresentation.

For demonstration purposes, let us assume that Meta's control effectiveness claims are valid. For the residual risk of misinformation, disinformation and hate speech to mathematically maintain Tier 1 and Tier 2 levels, some combination of these inferences must also be true:

- Residual impacts are confined to Meta's systems or individual users without broader ramifications to democratic processes and social stability (Scale rubric);
- Residual risks will not impact user's physical and psychological well-being and/or fundamental rights, but merely their general user experience (Nature rubric);
- Impacts will not affect society at large or vulnerable populations, but rather groups of repeated offenders (Demographic rubric);
- Fewer than one million users will likely be impacted by residual risks (Scope rubric); and
- Problematic occurrences are rare and will be seen if deliberately searched, rather than persistent with possibility of being found inadvertently (Volume rubric).

These are far from being self-evident affirmations and Meta's report presents no empirical data to support them across any of the Problems Areas assessed. These structural omissions render Meta's findings fundamentally unverifiable and potentially misleading. Omissions are particularly concerning when assessment depends on subjective and qualitative judgments, like the ones that comprise CSE% scores, questioning design robustness and application consistency of controls.

When unaccompanied by the necessary indication of criteria responses and empirical validation data, this self-referential approach further aggravates the industry's inherent vulnerability to bias. More than exposing Meta's failures, this analysis reveals how the DSA's self-assessment regime encourages selective reporting and systemic distortions by an industry whose interests directly conflict with those of regulators and the public. Without a fundamental redirection, such frameworks risk doing little more than institutionalizing performative compliance rather than enabling genuine accountability.¹²³

[A risk mitigation regime that replicates misplaced faith in the regulated industry](#)

By requiring platforms to make algorithmic adjustments to mitigate the systemic risks identified in their *own* risk assessments (Article 35), the DSA paradoxically replicates the very model of self-governance it intends to regulate, perpetuating a misplaced faith in self-correction. Rather than establishing clear architectural restrictions, or imposing independent oversight mechanisms over design, the provision presumes platforms will objectively assess and remedy risks created by their own systems, while remaining dependent on the very self-assessment processes that systematically understate harm. This circular logic institutionalizes unreliable compliance, as platforms hold both economic incentives to minimize risk findings and operational control over their measurement.

In requiring platforms to adjust recommender systems to mitigate systemic risks without prescribing real guidelines, benchmarks, design requirements or effective oversight mechanisms, the DSA fails to promote substantive accountability, reflecting a

¹²³ Cohen (n 40) 35.

trending regulatory preference for procedural compliance, which has thus far proven ineffective in curbing algorithmic harms.

Ultimately, Article 35 represents a missed opportunity to disrupt the recursive feedback loops that amplify societal harms by instituting measurable compliance standards for algorithm design and meaningful enforcement mechanisms.¹²⁴ Instead, it delegates crucial public interest protections to the very commercial entities whose operational models generate the systemic risks requiring mitigation, representing a fundamental conflict of interest that undermines the provision's effectiveness.

This regime also ignores the fact that algorithmic tweaks often have unintended harmful consequences.¹²⁵ For example, Facebook's 2018 shift backfired by amplifying misinformation, divisive content, and even real-world violence.¹²⁶ As Narayanan explains, we currently lack the methodological tools to verify whether tweaks will produce their intended effects and/or assess the long-term impacts of systemic risk accumulation. Even if platforms could be trusted to conduct tweaks intended to promote societal good, there simply "isn't a good way to measurably connect this kind of micro-level decision to its long-term impact."¹²⁷

An opt-out mandate that maintains *status quo* on consumer profiling

Finally, the opt-out provision for profiling encapsulated in Article 38(1)(c) of the DSA does little to disrupt status quo, seeing as platforms already provide such options. As Woodrow Hartzog and other scholars have pointed out, such rules operate on the false premise that users can and will make meaningful choices about complex data systems.¹²⁸ This assumption ignores the power asymmetries that make user choice illusory: platform architectures are inherently designed to manipulate people into agreeing to harmful practices. Personalization gratifies users, delivering them content that is both engaging and addictive, to the extent that they come to accept it as benevolent, a theoretical issue scholars refer to as 'psychological hedonism.'¹²⁹ At this

¹²⁴ Matias and Wright (n 17).

¹²⁵ Narayanan (n 31) 32.

¹²⁶ Hagey and Horwitz (n 112).

¹²⁷ Narayanan (n 31) 18.

¹²⁸ Hartzog (n 98) 631.

¹²⁹ Michal S. Gal, 'Algorithmic Challenges to Autonomous Choice' (2018) 25 *Michigan Technology Law Review* 59, 78 <https://repository.law.umich.edu/mtlr/vol25/iss1/3> accessed 15 August 2025.

point, users are unable, or unwilling, to break the cycle. To avoid this, regulation should ensure baseline protections no matter what users choose,¹³⁰ rather than relying on easily undermined opt-outs.

Additionally, as Cohen explains, individualized solutions, like opt-outs, are inherently infective against “predesigned, networked assemblages for data-driven patterning that operate at scale,”¹³¹ hence remaining inappropriate tools through which to address the normative challenge this study examines.

Chapter 3 - The DMA: A Missed Opportunity to Reshape the Market Dynamics that Underline Content Distribution

Enacted alongside the DSA, the Digital Markets Act (DMA) completes the EU's framework on digital regulation by addressing platform dominance from a market perspective. With the goal of promoting fairness and contestability in digital markets,¹³² it defines obligations directed at ‘gatekeepers’ who provide ‘core platform services,’ a definition that includes all major social media platforms, as per Article 2(2) and Article 3(1).

A key aspect addressed in the regulation is gatekeeper control over ‘rankings,’ a term broadly defined in Article 2(22) as encompassing all forms of algorithmic prioritization. The DMA subjects these systems to three core obligations: non-discrimination, with Article 6(5) prohibiting self-preferencing and unequal treatment of users; transparency, with Article 6(11) requiring ranking parameters be disclosed to both users and competitors, and independent oversight, with Article 15 mandating gatekeepers submit yearly audits of their consumer profiling techniques to the European Commission (EC). Collectively, these provisions aim to disrupt the black-box nature of ranking systems while creating enforceable accountability mechanisms.

The DMA's stated purpose of regulating gatekeeper power, coupled with its emphasis on promoting non-discriminatory treatment in rankings, suggest a potential to

¹³⁰ Hartzog (n 98) 631.

¹³¹ Cohen (n 40) 12.

¹³² European Commission, ‘The Digital Markets Act is the EU’s Law to Make the Markets in the Digital Sector Fairer and More Contestable’ (14 February 2024) <https://digital-markets-act.ec.europa.eu/index_en> accessed 15 August 2025.

meaningfully impact the systems governing content distribution on social media. However, this expectation is ultimately undermined by the Act's narrow commercial orientation, which focuses solely on the economic effects of platform control and discrimination while overlooking the broader societal impacts of algorithmic ranking—such as the suppression of marginalized voices and the amplification of misinformation and hate speech.

This limited commercial focus—apparent both in the framework's legislative history¹³³ and its early enforcement actions—¹³⁴ is illustrated by Article 6(5), which requires gatekeepers to apply “transparent, fair and non-discriminatory conditions” to the ranking of “services and products.” Although this obligation technically applies to social media, the framing around market goods effectively precludes meaningful application to the systems employed to rank and organize expressive content.

Alternatively, Article 15—which requires gatekeepers submit independently audited descriptions of their consumer profiling techniques to the EC—emerges as a potentially relevant mechanism to foster platform accountability. As defined in Article 4(4) of the General Data Protection Regulation (GDPR), profiling refers to any form of automated processing of personal data used to evaluate or predict aspects of a natural person's behaviour, preferences, or interests. This definition directly aligns with the predictive analytics that underpin content distribution in social media. Moreover, by requiring that audit reports be conducted by independent bodies, the DMA moves beyond the limitations of the DSA's self-assessment regime, offering the prospect of more effective oversight.

However, once again, a closer look at the disclosures actually submitted by the industry reveals how easily procedural oversight can be manipulated to create the appearance

¹³³ European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act)’ (Explanatory Memorandum, COM(2020) 842 final, 15 December 2020) <<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0842>> accessed 15 August 2025.

¹³⁴ Hein Hobbelen and David Wouters, ‘First-ever DMA Non-Compliance Fines – Another Step in Big Tech Scrutiny.’ (Competition Law Insights, 20 May 2024) <<https://competitionlawinsights.twobirds.com/post/102kajb/first-ever-dma-non-compliance-fines-another-step-in-big-tech-scrutiny>> accessed 15 August 2025.

of compliance while concealing underlying practices, highlighting the inherent limitations of these mandates as effective regulatory tools.

The DMA-mandated report published by Meta in 2025¹³⁵ is a ten-page document that largely consolidates descriptions already available in its Transparency Centre concerning ranking and data protection protocols, offering little in the way of additional detail. The report dedicates considerable space to arguing the benefits of consumer profiling, highlighting the legal bases for such practices under the GDPR and underlining in-product user controls (e.g., options to show more or less of certain content, hide posts, snooze suggestions, or unfollow accounts). Beyond this, the report offers only vague descriptions of the general logic of profiling and personalization, presented in self-congratulatory terms that highlight supposed benefits while neglecting the normative challenges at the centre of public debate.

Overall, there is no meaningful examination of whether Meta's machine learning models incorporate safeguards against discriminatory outcomes, or how the company measures and mitigates potential harms arising from its engagement-driven systems. Furthermore, the report's reliance on previously published material raises questions about whether it fulfils the DMA's likely intended purpose of maximizing observability of platforms' profiling practices.

Like the DSA, the DMA mandates reporting but lacks standardized methodologies and criteria for its rigor—allowing gatekeepers to treat this as a box-ticking exercise. Rather than diagnostic tools for public accountability, transparency mandates are once again appropriated as means for platforms to perform compliance while resisting to admit their role in perpetuating systemic harms. By disclosing just enough to satisfy the letter of the DMA while avoiding more contentious aspects of its profiling practices, Meta's audit report highlights both the limitations of the current regulatory framework and the need for more robust, preemptive mechanisms to counteract entrenched industry biases.

¹³⁵ Meta, *Meta Consumer Profiling Techniques: Digital Markets Act Non-Confidential Public Overview* (6 March 2025) <<https://transparency.meta.com/sr/meta-consumer-profiling-2025>> accessed 15 August 2025.

Chapter 4 - From Failure to Insight: What Can We Learn from Current Regulatory Standards

Policymakers worldwide are increasingly confronted with evidence that the most harmful pathologies in social media arise from the dangerous interplay of content distribution, surveillance, and profit-driven predictive targeting that constitutes platform business model.¹³⁶ Yet, despite the EU's global reputation as a leader in technology regulation, it has yet to implement mechanisms capable of addressing the democratic deficits this model creates. This gap is starkly evident in the failure to protect free speech and public discourse from the distortions of platform capitalism.

From the AI Act to the DSA and the DMA, EU regulation lacks a specific regime for such protections. Whenever content distribution is addressed, focus is either narrowly commercial (under the DMA's rankings regime) or ill-defined (under the DSA's recommender system regime). Moreover, responses remain procedural and reactive, constrained by the increasingly criticised assumption that imposing compliance will somehow reform fundamentally harmful practices.¹³⁷

Why have limitations proven so persistent? The preceding chapters have pointed to some of the practical and theoretical obstacles hindering more robust protections. In this chapter, we will analyse these barriers in depth, distilling them into three categories. By tracing some of the root causes of these failures—from theoretical misconceptions to practical challenges—we aim to illuminate a path that might allow more intentional and effective frameworks for supporting free speech and democratic participation.

Industry-captured myths and standards.

When examining the trajectory for meaningful AI regulation, Hartzog posits the indispensability of three elements: a strong administrative state, a humancentric conception of free expression, and the political will to enforce existing frameworks.¹³⁸ The actual course of digital regulation, however, reveals a starkly

¹³⁶ Cohen (n 40) 35.

¹³⁷ For example: Cohen (n 40); Riemer and Peter (n 8); Langvardt (n 3). Procedural mandates, particularly those directed towards transparency, are also criticised by computer scientists, see: Narayanan (n 31); Matias and Wright (n 17).

¹³⁸ Hartzog (n 98) 630.

different reality; marked by fragmentation, inconsistency, and undeniable co-optation by industry lobbying. In the first half of 2025, leading tech firms—including Meta, Alphabet, Microsoft, X, and OpenAI—collectively spent a staggering \$36 million on federal lobbying in the United States.¹³⁹ This aggressive spending mirrors their efforts in Europe, where, in 2023, the same companies spent over €113 million to influence digital policy.¹⁴⁰ This financial pressure continued in 2024, when Meta and Microsoft invested €9 million and €7 million, respectively, in EU lobbying.¹⁴¹

Internet industries would hardly made such investments in vain. Studies suggest that corporate capture has indeed diluted regulatory ambition on AI and platform governance, contaminating policy responses with questionable narratives designed to limit the scope and transformative potential of meaningful governance.¹⁴² Some of the common corporate notions that have influenced regulation include:

1. **Innovation fetishization**, representing the assumption that innovation is inherently beneficial to society, and that policymakers should encourage rather than constrain it—even as mounting evidence demonstrates its individual or social costs.
2. **Technological neutrality**, portraying the belief that technology is neutral and that risks derive mostly from misuse; therefore, regulation should target use, not the tools themselves; and
3. **Algorithmic opacity**, comprising the claim that algorithms are too complex to be technically understood by anyone other than their operators, thereby justifying

¹³⁹ Amelia Minkin, 'As Washington Debates Major Tech and AI Policy Changes, Big Tech's Lobbying is Relentless.' (*Issue One*, 22 July 2025) <https://issueone.org/articles/as-washington-debates-major-tech-and-ai-policy-changes-big-techs-lobbying-is-relentless/?utm_source=chatgpt.com> accessed 15 August 2025.

¹⁴⁰ Mared Gwyn Jones, 'Tech Companies Spend More Than €100 Million a Year on EU Digital Lobbying' (*EuroNews*, 11 of September 2023) <https://www.euronews.com/my-europe/2023/09/11/tech-companies-spend-more-than-100-million-a-year-on-eu-digital-lobbying?utm_source=chatgpt.com> accessed 15 August 2025.

¹⁴¹ Paula Soler, 'Big Tech, Banking, Energy: Who Are the Biggest Spenders on EU Lobbying?' (*EuroNews*, 24 of February 2025) <https://www.euronews.com/my-europe/2025/02/24/big-tech-banking-energy-who-are-the-biggest-spenders-on-eu-lobbying?utm_source=chatgpt.com> accessed 14 August 2025.

¹⁴² Kevin Wei and others, 'How Do AI Companies "Fine-Tune" Policy? Examining Regulatory Capture in AI Governance' (*Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 16 October 2024) <<https://doi.org/10.48550/arXiv.2410.13042>> accessed 15 August 2025.

that external oversight be limited to procedural mandates and reliant on self-assessments.

4. **Regulatory restraint**, constituting the idea that external oversight should never challenge the underlying incentive structures of internet industries, because doing so would supposedly stifle liberty and economic growth.

In her work on surveillance capitalism, Cohen explains how the dogma that technology is inherently collectively desirable has stifled regulatory responses to systemic threats.¹⁴³ Mislead by this assumption—and the notion that side effects must be accepted as inevitable costs of progress—regulators have produced reactive frameworks that prioritize risk mitigation over harm prevention, and procedural compliance over meaningful oversight and reform. The belief that technology advances ‘for our benefit’—and hence should be celebrated, rather than constrained—is, however, a myth.

While new AI tools might benefit us, they are not created *for* our benefit but for monetary gain. Consequently, as Hartzog explains, people can benefit from these technologies “and still be individually and collectively worse off overall.”¹⁴⁴ Ensuring AI systems are deployed for sustained societal gain is not a given. On the contrary, it requires active pursuit, notably through the institution of safeguards that can effectively protect our fundamental rights and democratic achievements from systematic erosion by the self-dealings of the companies that provide them.

From superficial transparency mandates to pro-forma self-assessments, policymakers have repeatedly deferred to corporate interests under the guise of ‘innovation-friendly’ governance. Future interventions must break this cycle by ensuring rules are designed for public benefit, not corporate convenience. In the realm of content distribution, this requires regulation that directly addresses the underlying incentive structures distorting public discourse and is intentionally designed to resist the commodification of free expression and democratic participation—something current frameworks have yet to

¹⁴³ “As we saw in Chapter 1, the financial transactions that produced the economic bubble of the 2000s and the ensuing global financial collapse were triumphs of complex information processing. Before and after the crash, both financial leaders seeking to avoid regulation and the government officials charged with oversight responsibility invoked ‘innovation’ to justify regulatory restraint.” In: Cohen (n 9) 92.

¹⁴⁴ Hartzog (n 98) 644.

achieve. The path forward requires rejecting the false supremacy of innovation over fundamental rights and reaffirming the essential role of regulation in achieving a balance that sustains both technological progress and fundamental rights protection. Liberal democracies have, after all, long relied on enduring mechanisms of constitutional balancing to navigate tensions between competing values—most notably through the principle of proportionality.¹⁴⁵ Why should now be any different?

Another myth widely endorsed by internet industries is that AI systems are neither good nor bad, but inherently amoral tools that can produce negative consequences when misused. This framing, reinforced by the fetishization of innovation, suggests that regulation should only target a tool's application, never the tool itself—an orientation that has become the hallmark of digital regulation. While the myth of technological neutrality myth has rendered the imposition of rules and restrictions on algorithmic design a persistent political taboo, it is increasingly criticised in legal and computational scholarship.

To pretend AI is neutral is “to allow companies to launder their moral choices that affect billions of people into machines and avoid responsibility for the reality they helped bring about.”¹⁴⁶ Whether directly coded by humans or self-learning, algorithms “develop their logic on the basis of human-controlled and -designed processes.”¹⁴⁷ They are neither objective nor neutral but “outcomes of human deliberation and power struggles.”¹⁴⁸ By refusing to scrutinize algorithmic design, regulators are in practice allowing private companies to be the sole voice of human deliberation and the sole arbitrators of our power struggles. Developing regulation that transcends the limitations of current frameworks and addresses both algorithmic design and the underlying incentives of

¹⁴⁵ “The proportionality/justice kinship can be interpreted as revealing an important aspect of the 'empirical adequacy' requirement for policy: proportionate regulatory responses to facts are assumed to be most apt to deliver justice.” In: Ljupcho Grozdanovski and Jérôme De Cooman, ‘Forget the Facts, Aim for the Rights! On the Obsolescence of Empirical Knowledge in Defining the Risk/Rights-Based Approach to AI Regulation in the European Union.’ (2023) 49 *Rutgers Computer & Technology Law Journal* 208, 321 <<https://orbi.uliege.be/bitstream/2268/307355/1/Grozdanovski%20and%20De%20Cooman%2C%20For%20the%20Facts%2C%20Aim%20for%20the%20Rights%20-%2049.2%20RutgersComputerTechLJ.pdf>> accessed 15 August 2025.

¹⁴⁶ Hartzog (n 98) 636.

¹⁴⁷ AlgorithmWatch, ‘Busted Internet Myth: Algorithms Are Always Neutral’ (25 November 2019) <<https://algorithmwatch.org/en/busted-internet-myth-algorithms-are-always-neutral/>> accessed 15 August 2025.

¹⁴⁸ Ibid.

platforms requires challenging the fallacy of algorithmic amorality in policymaking. Without concerted political and academic efforts in this direction, regulatory responses are likely to remain procedural and overly deferential to corporate interests, as recent EU experience has demonstrated.

Digital regulation in the EU, and worldwide, has also been largely influenced by the assumption that algorithms are too complex to be meaningfully overseen. This belief is reinforced by the industry in DSA and DMA-mandated reports. For example, by describing its systems as employing “hundreds” of predictive models, each informed by “thousands” of dynamic input signals,¹⁴⁹ Meta implies technical complexity beyond state scrutability. By perpetuating this notion, platforms like Meta position themselves as the sole entities capable of understanding their own systems, thereby forcing regulators and users to rely on self-reporting and internal audits.

Like the myths of algorithmic neutrality and the fetishization of innovation, this presumption must also be interrogated and dismantled if we are to pursue meaningful, democratic governance. While algorithms do indeed process massive quantities of data, it is a mistake to assume that this renders external oversight impossible. As computational scientist and Princeton University professor Arvind Narayanan has argued, understanding what recommendation algorithms are trained to do is not only possible, but conceptually straightforward. In his view, these systems are “like any other technology—say, a car or a smartphone.”¹⁵⁰ While many of their details may be complex or proprietary, we do not need to understand them at a granular level to know if they are working.

To safeguard free speech and democratic participation, effective oversight need not involve real-time monitoring of every data point or variable. Instead, it must ensure that algorithmic systems are designed—whether by intent or consequence—to amplify, rather than erode, democratic values. For example, regulators should be able to determine if platforms systematically suppress marginalized voices simply because they are deemed ‘less profitable’ and how that affects public discourse. Such assessments

¹⁴⁹ See footnote n 102.

¹⁵⁰ Narayanan (n 31) 6.

need not rely on exhaustive data analysis; techniques like comparative simulations and AI audits could suffice. As free speech theorist Alexandre Meiklejohn once argued, the essence of free speech is not that all may speak, but that “everything worth saying shall be said.”¹⁵¹ Similarly, the goal of free speech-oriented oversight should not be to scrutinize every individual act of ranking or amplification, but to ensure that algorithmic systems broadly uphold the conditions necessary for meaningful public discourse. This includes safeguarding the diversity of voices and promoting the equitable distribution of their reach.

Recognizing that the complexity and fluidity of algorithmic systems do not preclude oversight—and understanding what types of information are truly relevant for effective regulation (such as whether speech-distribution processes are intentionally or foreseeably structured in ways that undermine rather than support democratic values)—opens the door to more ambitious regulatory models. These models could move beyond superficial procedural compliance and instead confront the deeper logics of algorithmic design, deployment, and incentive structures.

Flowing logically from the myths discussed above is the entrenched belief that regulation should not interfere with the underlying incentive structures of tech companies, on the grounds that doing so would stifle liberty and innovation. This assumption is both historically and normatively flawed and ignores the very purpose of regulation: protecting public interest goals by imposing external constraints on economic activity.

Regulatory interventions in other sectors—such as environmental and consumer protection—¹⁵² have long required the reshaping of core incentive structures to drive systemic change. While these efforts have always faced significant resistance from regulated industries and liberals, their achievements have managed to deliver significant progress without imposing excessive burdens on individual liberty or stifling innovation. Such outcomes have been made possible through sustained political and academic engagement and, at a policy level, through enduring mechanisms of constitutional

¹⁵¹ Pallavi Guniganti, ‘Alexander Meiklejohn: Teacher and Citizen’ (2005) *Ideas of the First Amendment* (Thomson/West 2006) <<https://ssrn.com/abstract=2104006>> accessed 23 May 2024

¹⁵² Cohen (n 9) 184.

balancing—most notably, the principle of proportionality, a foundational tool in liberal democracies.

Despite the vast resources invested in lobbying, digital platforms have yet to offer a convincing rationale for why they should stand outside the constitutional frameworks that constrain every other sector. The fact that they have managed to do so for so long underscores the urgency of empowering regulators to critically evaluate algorithmic outcomes through principles such as proportionality and, if necessary, impose reforms on the economic logic that drives them. To neglect this task is not merely to defer to corporate self-governance; it is to abdicate the constitutional and democratic duty of regulation itself.

The Technology Normalization Cycle

As previously outlined, the provisions that address content distribution in EU regulation (encapsulated mostly in the DSA) have so far been constricted in a reactive, procedural, risk-mitigation, mindset borrowed from traditional domains like health and consumer safety.¹⁵³ These models, designed to address physical harms, are generally ill-suited to confront the systemic threats posed by digital systems,¹⁵⁴ which are far more subtle and pervasive. The ‘technology normalization cycle’, examined by Hartzog, explains why.

From initial public critique through gradual acclimation to eventual institutional acceptance, the implementation of invasive technologies follows a predictable pattern.¹⁵⁵ When faced with a new ethically questionable tool or feature, users’ initial reaction may be one of suspension or rejection. However, eventually, repeated exposure eases them into habituation, dulling collective resistance into a feeling that “this is our new reality, so you might as well get used to it.”¹⁵⁶ The longer a tool is deployed, the less questionable it becomes. As it fades into the background, people begin to view it as acceptable, or even desirable, reflecting the psychological bias of ‘favourably disposed normalization.’¹⁵⁷ Lawmakers then codify this surrender through superficial procedural hurdles that create the appearance of oversight while leaving the

¹⁵³ Grozdanovski and Cooman (n 145) 295.

¹⁵⁴ Ebers (n 91) 7.

¹⁵⁵ Hartzog (n 98) 612.

¹⁵⁶ Ibid.

¹⁵⁷ Ibid, 614.

systems themselves fundamentally intact. Soon, a new technology is put in place and the cycle restarts.

The systematic repetition of this cycle is eroding society's capacity to resist it.¹⁵⁸ As we became conditioned to accept surveillance and exploitation as inevitable features of modern life, we tolerate ever more invasive practices. When citizens grow so powerless they can no longer imagine rejecting harmful technologies, democratic self-governance becomes impossible.¹⁵⁹ Through gradual exposure and desensitization, people are manipulated into consenting to systems that ultimately undermine their own autonomy. At this point, reactive procedural measures—like transparency reports, self-assessments, and user controls—are meaningless because they are unable to mobilize a desensitized public or compel institutional action against entrenched harmful practices. Instead, they just further legitimize wrongdoing.¹⁶⁰

Meaningful reform requires increasing general awareness of this cycle and equipping regulators to respond to it before it is too late. This entails shifting from a risk-based approach, oriented towards compliance, to a rights-based approach, “focused on ‘how to guarantee fundamental rights and democracy.’”¹⁶¹ It also requires orienting regulation towards effectively preventing the risks of modern technology rather than merely mitigating the supposed ‘inherent costs’ of progress.

Industries that require regulation against systematic harms face overwhelming financial incentive to resist change in the first place. This inevitably leads them to exploit loopholes, flexibilities and trust-based obligation to protect profitable but harmful practices.¹⁶² The predictable outcome of an economic system that rewards harm is negligence. Co-regulatory schemes that allow excessive flexibility and self-policing (such as the ones developed in the DSA and the DMA) should, therefore, be viewed with scepticism. As demonstrated by previous analysis, Meta's DSA and DMA-mandated reports have demonstrated little more than the company's ability to produce auditing paperwork. In contrast, directing regulation towards clear outcomes—in this case,

¹⁵⁸ Ibid, 616.

¹⁵⁹ Hartzog (n 98) 617.

¹⁶⁰ Riemer and Peter (n 8) 422.

¹⁶¹ Grozdanovski and Cooman (n^o 145) 296.

¹⁶² Hartzog (n 98) 631.

upholding free speech and democratic participation—will enable regulators to assess a platform’s actual capacity to uphold standards that genuinely serve users and the broader public.

When platforms deliberately engineer user complacency by systematically eroding our capacity to resist harm, mitigation becomes futile. True protection demands dismantling the architectures that enable platforms to manipulate users’ and regulators’ risk tolerance in order to reestablish the general public’s ability to resist change that is not good for them. Although internet industries often claim precautionary policymaking would stifle liberty and economic growth, they have yet to provide compelling justification for evading scrutiny that has long been central to progress in other domains, where regulation has advanced both public welfare and industrial innovation without sacrificing liberty.¹⁶³

Recognizing that the commodification and exploitation of speech are not inevitable features of modern life, but rather engineered outcomes of a deliberate project to normalize extractive technologies and deflect regulatory scrutiny, is a crucial step towards meaningful reform. This understanding empowers regulators to break the cycle by moving beyond today’s limited, *ex-post* procedural compliance toward proactive regulatory frameworks focused on securing and promoting free speech and democratic participation.

The Bottleneck Effect of EU Regulation

As previously acknowledged, the AI Act's failure to include content distribution algorithms within its risk classification framework raises a critical concern as to whether this was a deliberate deregulatory choice or an oversight by EU regulators. If it was a deliberate choice, what political or technical considerations prevented even baseline restrictions on such technologies despite mounting evidence of their societal harms? EU law and AI specialists Ljupcho Grozdanovski and Jerome De Cooman provide an assessment of the fact-neutral approach underlying the AI Act’ methodology, offering insights that might help shed light onto these questions.

¹⁶³ Cohen (n 9) 184.

As they explain, “[a]n overview of fact-finding procedures in the field of AI shows that, when choosing which risks to address, regulators were generally disinterested in statistical evidence on the possibly harmful features of various systems. They seemed more interested in how representative groups felt about certain preselected risks.”¹⁶⁴ In other words, policy preferences—rather than empirical evidence—appear to have served as the ultimate relevance criterion in risk identification during the drafting process. Diverging from the traditional bottom-up approach where factual data guides policymaking, the regulatory choices encompassed in the AI Act are not based on empirical knowledge per se, but in axiological congruency “with the foundational values that underlie the overall EU approach toward regulating new technologies.”¹⁶⁵

This analysis suggests that the omission of content distribution from the AI Act may reflect the low prioritization of this application within the EU’s policy agenda. Whether due to political sensitivities or the influence of substantial corporate lobbying, this blind spot is not unique to the EU, finding a familiar parallel in Brazil’s ongoing efforts to regulate AI.

In June 2024, the Brazilian Senate consolidated multiple AI regulation proposals into a single unified bill. This draft explicitly classified “large-scale, significantly automated systems for content production, curation, dissemination, recommendation, and distribution [...] designed to maximize user engagement time”¹⁶⁶ as high-risk systems. Such designation would attract a series of obligations under the bill’s original wording, including record-keeping, automated monitoring to track potentially discriminatory outcomes, comprehensive testing on sector-specific safety standards, ensuring the statistical adequacy and contextual relevance of training data, establishing bias prevention measures, and ensuring decision explainability.

However, subsequent amendments, approved in December 2024, not only withdrew content distribution from the high-risk regime but eliminated any reference to these

¹⁶⁴ Grozdanovski and Cooman (n 145) 260.

¹⁶⁵ Ibid 285.

¹⁶⁶ Free translation from: Comissão Temporária Sobre Inteligência Artificial no Brasil, *Parecer s/n* (7 June 2024). https://legis.senado.leg.br/sdleg-getter/documento?dm=9630164&ts=1742240894034&rendition_principal=S&disposition=inline accessed 15 August 2025.

systems entirely.¹⁶⁷ Only five days later, the Brazilian Senate approved the resulting draft, which is currently under deliberation in the House of Representatives.¹⁶⁸ The current draft of the Brazilian AI Bill replicates the AI Act's omission, leaving platforms' speech-governing powers virtually unchecked. Even more strikingly, this suppression occurred without any accompanying justification, disregarding standard legislative practice that substantive changes be supported by clear and reasoned explanation.

Although repercussion remains timid, the Senate's decision has faced some criticism, with Bia Barbosa, advocacy coordinator for the organization Reporters Without Borders, evaluating that "[i]t makes no sense to have an AI regulation bill that does not address recommendation and content moderation systems, which are high-risk systems. However, platforms—just as they do in every country in the world—strongly oppose any regulation that could affect their businesses (...)."¹⁶⁹

Drawing a parallel between the abrupt removal of the regime intended to regulate content distribution in Brazil and the AI Act's similar omission is inevitable. As policymakers increasingly look to the EU as a sort of global leader in matters of digital regulation, its frameworks both shape and constrain instruments worldwide. This is problematic because the EU's regulatory architecture is grounded in constitutional limitations that, while not universally relevant,¹⁷⁰ end up functioning as an unintentional ceiling on regulatory ambition.

On one hand, the Treaty on the Functioning of the European Union (TFEU) requires legislators to act only where sector-specific harmonization is necessary (Articles 3–6). This restriction has prevented the development of a comprehensive digital governance strategy, leading to fragmented and sometimes incoherent regimes.¹⁷¹ The fact that the

¹⁶⁷ Comissão Temporária Sobre Inteligência Artificial no Brasil, *Parecer 1/2024* (5 December 2024) <https://legis.senado.leg.br/sdleg-getter/documento?dm=9858756&ts=1738768138734&rendition_principal=S&disposition=inline> accessed 15 August 2025.

¹⁶⁸ Comissão Temporária Sobre Inteligência Artificial no Brasil, *Parecer 208/2024* (10 December 2024) <<https://legis.senado.leg.br/sdleg-getter/documento?dm=9884955&ts=1742240908365&disposition=inline>> accessed 15 August 2025.

¹⁶⁹ Free translation from: Lucas Pordeus León, 'Big Tech é excluída do rol de Inteligência Artificial de alto risco' (*Agência Brasil*, 3 December 2024) <<https://agenciabrasil.ebc.com.br/politica/noticia/2024-12/big-tech-e-excluida-do-rol-de-inteligencia-artificial-de-alto-risco>> accessed 15 August 2025.

¹⁷⁰ Ebers (n 91) 20.

¹⁷¹ Ebers (n 91) 14

AI Act excluded content distribution from its risk classification even after the DSA had already recognized the systemic harms of these systems, illustrates this dissonance and the problematic of indiscriminately importing EU frameworks to other jurisdictions.

On the other hand, as AI Act critic Martin Ebers has observed, the lack of general competence to harmonise fundamental rights has led the EU to ground its digital regulation framework on the internal market competence conferred by Article 114 TFEU.¹⁷² This limitation structurally constrains EU frameworks to a risk-based approach that prioritizes product safety and risk management while neglecting “the minimum and non-negotiable nature of human rights.”¹⁷³ Even when frameworks somewhat manage to deviate from a strict rights-based approach by incorporating elements of fundamental rights protection,¹⁷⁴ scholars argue the result “merely gives idiosyncratic expression” to ECFR-enshrined rights, assuming appropriate levels of protection will eventually be achieved through standardization.¹⁷⁵ Such narrow, market-centric models are ill-suited for worldwide application. Yet, EU digital policy frameworks are still widely regarded as universal blueprints,¹⁷⁶ frequently imported without the critical assessment and adaptation they intrinsically require.

The limitations of the DSA, the DMA and the AI Act which we have previously explored are, therefore, not merely a result of political bias or theoretical misconceptions; they are also products of structural constraints embedded in the EU’s very constitution. As we continue to embrace the EU as the ultimate political leader in matters of digital policy, these structural constraints have channelled regulatory approaches that are unnecessarily narrow outside the EU jurisdiction. Nation-states—both in and outside the EU—retain the constitutional authority to pursue unconditional protections for fundamental rights and democratic values. This institutional capacity presents not only an opportunity but an obligation to develop alternative models that are better equipped to fulfil nation-states’ constitutional mandate of upholding fundamental rights and

¹⁷² Ibid 20

¹⁷³ Ibid 7.

¹⁷⁴ Grozdanovski and Cooman (n 145) 330.

¹⁷⁵ Ibid.

¹⁷⁶ Cristiano Codagnone and Linda Weig, ‘Leading the Charge on Digital Regulation: The More, the Better, or Policy Bubble?’ (2023) 2 *DISO* 1, 4 <<https://doi.org/10.1007/s44206-023-00033-7>> accessed 15 August 2025.

democracy. While EU frameworks on digital policy remain influential, their structural limitations advice policymakers elsewhere regard it as a floor, rather than a ceiling, for fundamental right protection.

Part III: The Solution: A Roadmap for Intentional Reform

Chapter 1 – What is Free Speech and How to Protect it in the Digital Age: Constructing a Theoretical Foundation for Reform

The propagation of myths and standards that reflect industry interests, the deliberate normalization of harmful practices, and the bottleneck effect created by the EU's constitutional constraints are some of the key factors that have influenced digital policy, preventing more meaningful interventions on this study's subject matter. Overcoming these challenges requires more than new regulation, it requires a new way of thinking about algorithmic governance that privileges people, democratic values and free expression before profits. Establishing the theoretical foundations for such reform is not merely an academic exercise, it is a necessary condition for change.

Experience with regulatory strives in other fields, like environmental protection and pharmaceutical regulation,¹⁷⁷ has demonstrated that confronting market power and entrenched institutional inertia requires sustained academic and political commitment. Transitioning from today's procedural paradigm to a proactive, rights-based response on the challenges of algorithmic content distribution will similarly require a coherent conceptual vision of what it is that we value—why free speech and democratic participation are important—and how best to preserve that which we value amidst economic, technological and sociological change. Without such understanding, interventions risk remaining reactive, fragmented, and ultimately ineffective.

As a leading scholar investigating how the digital age has reshaped fundamental rights theory—particularly the right to freedom of speech—Jack M. Balkin offers a compelling framework for grounding such understanding. While his theory is nuanced and

¹⁷⁷ Cohen (n 9) 184; Amy Kapczynski, 'The Law of Informational Capitalism' (2020) 129 *Yale Law Journal* 1460, 1491 <https://www.yalelawjournal.org/pdf/KapczynskiBookReview_iqh4qxtw.pdf> accessed 15 August 2025.

multifaceted, for the purposes of this investigation it can be productively approached through four core ideas.

Free Speech is an Infrastructure for Democratic Culture

Throughout his work, Balkin argues that digital technologies have fundamentally transformed the social dynamics of communications and speech.¹⁷⁸ By bringing into sharper focus five characteristics of expression that have always existed—but are now inescapable—the internet has exposed the limitations of twentieth-century models that reduced free expression to an instrument of democratic deliberation.¹⁷⁹ What are these five characteristics, or values? First, the digital sphere has demonstrated that free speech encompasses a vast array of subjects, extending beyond politics to include popular culture and everyday discourse. Second, it showcases that ordinary individuals are capable of remarkable innovation and creativity, disrupting traditional hierarchies of cultural production. Third, it has revealed that this creativity operates collaboratively, often advancing through adaptation and appropriation of existing ideas. Fourth, digital speech is interactive, it merges reading with writing, and production with consumption. Finally, it reminds us that it is through speech, that ordinary individuals have the chance to “participate in the growth and development of the cultures and subcultures that, in turn, help constitute them as individuals.”¹⁸⁰

By evidencing freedom of speech’s undeniably populist nature of, “its creativity, its interactivity, its importance for community and self-formation,”¹⁸¹ the internet has laid bare its true function and value. Contrary to what twentieth-century thinkers assumed, it is not merely concerned with enabling democratic deliberation, but with allowing individuals to create, interact, and participate freely in cultural production.¹⁸² In Balkin’s conception, free speech is an essential infrastructure for democratic culture; the foundation through which democratic participation and cultural production take shape. Culture, Balkin contends, is simultaneously an ingredient of society and of the self.¹⁸³

¹⁷⁸ Balkin (n 6); Balkin (n 19) and Balkin (n 65).

¹⁷⁹ Balkin (n 6) 32.

¹⁸⁰ Ibid.

¹⁸¹ Ibid.

¹⁸² Ibid.

¹⁸³ Ibid 37.

Freedom of speech, therefore, entails the freedom to participate in the narratives and forces that shape society and the self: democracy, in its truest form.

At its core, Balkin's framework redefines free speech as a mechanism for empowering ordinary individuals to participate in the practices and meaning-making processes that shape their identities and social realities, allowing them to influence the cultural forces that shape their lives.¹⁸⁴ Under this re-conception, protecting free speech requires more than safeguarding individual formal rights to expression, it requires ensuring that individuals have a genuine and fair opportunity to participate in shaping the narratives that define their identities, communities, and societies. This entails: (1) the right to publish, distribute, and actually reach an audience; (2) the right to interact meaningfully with others, influence and be influenced through cultural transmission and intellectual discourse; (3) the right to creative appropriation and innovation, permitting individuals to remix, annotate, and transform existing cultural materials into new expressions shared with others, and (4) the right to participate in the development of the communicative and cultural frameworks that constitute individual and collective identity.¹⁸⁵

The disconnect between Balkin's democratic culture ideal and the realities of platform governance reveals the contour of a normative crisis, which we have already described. When social media companies claim to enable democratic discourse, but their architectures subject our practical ability speak—*i.e.*: reach audiences, influence and be influenced, and participate in the development of cultural frameworks—to profit-driven imperatives, their legitimacy as stewards of free expression must be called into question. As previously analysed, a governance model that is primordially concerned with the commercial viability of expression and with maximizing engagement at all costs will systematically constrain opportunities for substantive dialogue and corrupt the mutual processes of cultural influence that sustain democracy.¹⁸⁶ Against this capitalist reduction of expression to commodity logic, Balkin's democratic culture model insists on equitable participation in meaning-making as a fundamental free speech right.

¹⁸⁴ Ibid 33.

¹⁸⁵ Ibid 43.

¹⁸⁶ See: Langvardt, (n3); Cohen (n 9); Cohen (n 40); Riemer and Peter (n 8); Balkin (n 6) and Balkin (n 19).

This framework emerges as an urgent counterpoint to the capitalist instrumentalization of expression that has become the hallmark of social media and reasserts communication as a collective good rather than a corporate commodity. Where capitalism dictates public discourse be shaped for monetary extraction, a model based on the ideal of democratic culture demands that communication networks preserve the necessary conditions for authentic cultural production, ensuring our political economy does not subjugate democratic participation.

Platforms Perform Essential Public Functions and Should be Regulated as Such

More than proposing an updated, democratic ideal of free speech to counter social media's commodification logic, Balkin also provides a constructive path for regulation. Like other scholars,¹⁸⁷ he believes that, although their infrastructures may be privately owned, communication networks perform functions that are essentially public in nature. As a result, they should be intentionally regulated to serve public interest. Without such regulation, Balkin warns, “powerful private interests will trample on free speech values in the relentless pursuit of profits.”¹⁸⁸ This profit-driven model has already created significant democratic deficits, necessitating urgent correction.

Acknowledging that the business model of social media entails “regular and pervasive content-based decisions,”¹⁸⁹ Balkin frames these platforms as ‘content curators.’ In his view, these entities have evolved into important custodians of the public sphere and of democratic self-government. Much like traditional mass media in the twentieth century, they now function as essential conduits for democratic participation, enabling engagement with art, politics, and culture. They structure and mediate public discourse, making collective conversations navigable and accessible. More critically, they also actively shape public opinion through algorithmic curation of content, personalized information feeds, and the enforcement of community standards.¹⁹⁰

By virtue of these unique roles, Balkin contends social media platforms bear significant moral responsibilities. These include not only allowing users to speak, but also actively

¹⁸⁷ Particularly, Cohen (n 40) 40.

¹⁸⁸ Balkin (n 6) 22

¹⁸⁹ Balkin (n 19) 2038.

¹⁹⁰ Ibid 2041.

maintaining the structural conditions that enable speech to matter—ensuring all participants can meaningfully contribute to the ongoing creation of their cultural environment.¹⁹¹ As Balkin emphasizes, this requires platforms to refrain from creating artificial barriers to access, discriminatory gatekeeping mechanisms, or hierarchical content privileging that distorts cultural flows—¹⁹² which once again conflicts with the realities of platform governance.

When applied to the concerns of this study, Balkin’s perspective on platform regulation helps inform specific policy interventions aimed at the incentive structures that distort public discourse, cultural production, and democratic participation. By acknowledging that platforms have an inherent moral duty to ensure meaningful participatory opportunities for all users—and that the state should enforce this obligation through regulation—he opens a promising avenue for intervention that addresses root causes rather than the superficial symptoms of distortion, transcending the limitations of current frameworks. This potential becomes even clearer when Balkin’s ideas are brought into dialogue with other interdisciplinary approaches, as demonstrated in the following chapter (see Chapter 2).

Protecting Free Speech Through the Promotion of Free Speech Values

Moving one step further, Balkin also advances a vision of how regulation might protect his updated conception of free speech. This vision emerges crucially from a critical assessment of the insufficiencies of twentieth-century protections. Balkin contends that by altering our understanding of free speech from a tool for democratic deliberation to an infrastructure for cultural engagement, the digital age demands a shift in how that liberty might be adequately protected.

Throughout the twentieth century, the primary method of safeguarding freedom of speech relied on judicially enforceable individual rights, particularly constitutional rights.¹⁹³ However, this approach ensures only a formal liberty to speak, a guarantee that becomes increasingly less valuable as our communication technologies develop ever more efficient methods of prioritizing corporate interests over popular

¹⁹¹ Ibid.

¹⁹² Ibid.

¹⁹³ Balkin (n 6) 47.

participation.¹⁹⁴ When free speech is understood not merely as the sum of individual rights, but as an infrastructure from democratic culture, protecting it requires a corresponding shift to proactive infrastructural maintenance. Such upkeep entails preserving equitable access to audiences and non-discriminatory opportunities of participation, as well as promoting popular agency and democratic oversight of the designs of cultural tools.¹⁹⁵ Otherwise, we will be left with purely formal guarantees of speech “embedded in technologies of control that frustrate their practical exercise.”¹⁹⁶

Grounded on this understanding, Balkin suggests policymakers can improve the infrastructural conditions of expression on two fronts: first, by promoting policies that facilitate popular participation, and second, by imposing design mandates that decentralize control and facilitate popular participation in communication technologies.¹⁹⁷ The system of free expression, he explains, has always depended on participatory policies (i.e.: public education, public libraries, public mail system, government subsidies for books, and so on) and on democratic design mandates (i.e.: common carrier rules in telephony, public interest rules in telecommunications, etc.) to sustain a thriving democratic culture.¹⁹⁸ While these structural pillars have long operated in the background of judicial enforcement, the challenges of the digital age have revealed the insufficiency of individual free speech rights, demanding we elevate them to the foreground.

More specifically, Balkin proposes regulation should work to infuse policy and technological design with free speech *values*. What are these values? They reflect the very characteristics that have always constituted expression and that digital technologies have made more salient: interactivity, popular participation, equitable access to technology, democratic governance over technical architectures, and creative agency to adapt and repurpose technological and cultural resources.¹⁹⁹ Ultimately, this is where our deepest regulatory commitment must lie: in preserving the participatory essence of free speech itself. By anchoring our efforts here, we can reconcile platform

¹⁹⁴ Ibid 48.

¹⁹⁵ Ibid 49.

¹⁹⁶ Ibid 48.

¹⁹⁷ Ibid 49.

¹⁹⁸ Ibid.

¹⁹⁹ Ibid.

governance with democratic values and steer technological evolution to serve, rather than override, the democratic essence of free speech.

Preserving Free Speech Requires Adaptation

Underlying Balkin's framework is the idea that protecting free speech in the digital age requires fundamental adaptation. We can either acknowledge that technology has transformed what it means to have freedom of expression, and update our institutions accordingly, or we can watch the erosion of a stagnated version of our twentieth-century liberties. These arguments are an outgrowth of a more general way of thinking about rights and fundamental liberties, called dynamic theory of rights, or rights dynamism.²⁰⁰ This ideology acknowledges the shifting nature of fundamental rights and how they tend to take on new dimensions as they are invoked in new contexts and actors with different interests and agendas. The digital age, Balkin argues, is transforming the meaning of freedom of expression—in many ways, it already has. The critical question is whether we will take an active part in this change or let it be imposed upon us. After all, "[w]hat we do not do for ourselves will surely be done to us."²⁰¹

As a society, we are called to retrace our steps and reassess our priorities to understand what the free speech principle is about—what it has always been about, "but only now can be adequately expressed."²⁰² It's essential, yet newly salient, participatory values. As Balkin describes, the experience is not of making something new, but of "finding something old, of recognizing principles and commitments already dimly understood, which suddenly are thrown into sharper focus by the alteration in our circumstances."²⁰³ Through this exercise, we gain clarity of how to preserve free expression from continuous erosion under the pressures of platform capitalism and shifting social and economic conditions.

What seems like adaptation is, in fact, a necessary act of preservation.²⁰⁴ Free speech endures not through rigid doctrine, but through continual recalibration between its evolving conceptual forms and the core, non-negotiable principles that come into

²⁰⁰ Balkin (n 6) 51.

²⁰¹ Ibid 54.

²⁰² Ibid

²⁰³ Ibid.

²⁰⁴ Ibid.

sharper focus as new social and technological contexts reshape the conditions of expression. This capacity for adaptation is the critical strength of Balkin's framework. Although it requires a much deeper commitment, it also allows sustained preservation that might never be achieved through superficial procedural measures.

Chapter 2 - Strengthening Balkin's Framework Through Interdisciplinary Dialogue

Balkin's theory for protecting free speech in the digital age offers a vital framework for confronting the corrosive influence of capitalism on the structures of online expression and cultural production. More than a critique of technological disruption, it offers a constructive path forward that entails:

- (i) Conceiving free speech as an essential infrastructure for democratic participation—requiring infrastructural investment to thrive;
- (ii) Recognizing the unique public function performed by “content curators” in structuring discourse, facilitating cultural production, and arbitrating access to audience—and acknowledging that such function must be regulated to secure public interest;
- (iii) Empowering regulators to respond to the changing social and material conditions of expression by reaffirming and amplifying its foundational participatory values, and
- (iv) Acknowledging that the only way to preserve fundamental rights from erosion is through intentional adaptation.

Already powerful, this framework gains additional robustness when synthesized with the interdisciplinary scholarship that informs this study. In the field of law, Balkin's work resonates with the insights of other contemporary scholars that investigate structural solutions to the issues to platform capitalism. A particularly helpful intersection arises from Cohen's scholarship on ‘infrastructure thinking.’²⁰⁵ Conceived to describe technology as a system of power, this concept allows regulators to assess the underlying arrangements that enable and shape large-scale exchanges; to probe them “downward

²⁰⁵ Hartzog (n 98) 638.

and outward,”²⁰⁶ to identify and address the root causes of platform harm beyond surface-level symptoms. In her words: “infrastructure thinking aims to expose those arrangements and consider what they ask us to take for granted.”²⁰⁷

Such approach presents a compelling tool for diagnosing harm and a viable pathway to materialize the promotion of free speech values in policy and design. When applied to content distribution, Cohen’s expansive analytical lens reveals the dysfunctions of the digital public sphere—hate speech, misinformation and polarization, amongst others—as intended or unintended consequences of algorithmic monopoly and platform business model,²⁰⁸ offering a strong normative bases for intervention.

Focussing specifically on free speech, Karl Langvardt’s scholarship on the online public sphere also provides substantial contributions. The author underscores how algorithmic governance prioritizes polarization and distraction over deliberation.²⁰⁹ In addressing this issue, he explores structural alternatives that range from restraining advertising practices, to regulating platform’s “operational design in the hopes of creating an ‘environment that was less prone to polarization and toxic virality.’”²¹⁰ He also examines the promotion of external measures to counter platforms’ unprecedented control over information flows. Such solutions, which will be further developed in this study’s final chapter, tightly align with Balkin’s framework, providing crucial insights of how his vision for preserving free speech’s participatory essence might be achieved.

Presenting a unique framing that focuses on how engagement-driven designs sever the direct relationship between speakers and audiences, Riemer and Peter’s work on algorithmic audiencing also dialogues with Balkin’s call for regulation that restores democratic control over access to audience and visibility. Their stance that any attempt to regulate algorithmic governance without confronting the “economic logics of monetising speech-as-content will miss the mark,”²¹¹ is particularly insightful and further substantiates Cohen and Langvardt’s model for structural intervention.

²⁰⁶ Cohen (n 40) 17.

²⁰⁷ Ibid 17.

²⁰⁸ Ibid 12.

²⁰⁹ Langvardt (n 3) 2.

²¹⁰ Ibid 52/53.

²¹¹ Riemer and Peter (n 8) 410.

Although it focuses mostly on data governance, Hartzog’s perspective that regulation must address “key dynamics” to ensure that the risks of AI deployment do not override its benefits,²¹² also substantiates a model that advances interventions in platform and algorithmic design. The concrete measures Hartzog proposes will also be examined in the following and final chapter of this analysis.

Taken together, these perspectives both support Balkin’s vision for a proactive framework on free speech protection—one that focusses on promoting participatory polices and decentralizing control over technological design—and help develop it into a more concrete model that specifically contemplates the need to address platform business model and algorithmic design.

Outside of legal scholarship, this approach might find support in the works of Arvind Narayanan, renowned computational sciences and Princeton University professor, who attributes engagement-optimization as the primary cause of “[m]any of the familiar pathologies of social media.”²¹³ Although his focus is not to analyse the legality of platform practices, or to assess the insufficiency of current paradigms, by estimating improving algorithmic transparency will not contain platform-related harms,²¹⁴ Narayanan suggests support for more structural interventions.

In the fields of communication and political science, Zizi Papacharissi’s critical analysis of the failed promise of a ‘virtual public sphere’²¹⁵ is also relevant. While acknowledging the affordances of new technologies for broader participation, her work provides empirical support to the conclusion that such participation often reinforces existing hierarchies rather than diversifying democratic engagement.²¹⁶ Papacharissi’s observation that public discourse is often unequal, uncivil, and too amorphous to fully resolve democratic deficits²¹⁷ may, at first glance, seem to misalign with Balkin’s normative vision of free speech as an infrastructure for democratic culture. However,

²¹² Hartzog (n 98) 644.

²¹³ Narayanan (n 31) 35.

²¹⁴ Ibid.

²¹⁵ Zizi Papacharissi, ‘The virtual sphere The internet as a public sphere.’ (2002) 4 *New Media & Society* 9, 22 <<https://doi.org/10.1177/1461444022226244>> accessed 15 August 2025.

²¹⁶ Ibid.

²¹⁷ Ibid 11.

this tension is ultimately productive, seeing as her critique reinforces the urgency of redesigning the structural conditions of digital discourse to strengthen the conditions of democratic participation.

Finally, Balkin's approach, rooted in a dynamic conception of fundamental rights, aligns closely with the ideology of digital constitutionalism. This ideology seeks to adapt traditional constitutional principles—such as the rule of law, separation of powers, and fundamental rights protection—to the realities of the digital age.²¹⁸ It aims to 'constitutionalize' the virtual environment, ensuring that foundational democratic values persist despite technological transformation. Edourdo Celeste, an exponent of digital constitutionalism, explains that this requires reinterpreting analogue norms to address digital complexities, while preserving core human rights protections.²¹⁹ Although digital constitutionalism is not directly referenced in Balkin's work, its approach aligns closely with his vision of preservation through adaptation. Both approaches reject the notion of the digital sphere as a secluded world exempt from constitutional safeguards, instead affirming that it must remain subject to the same democratic principles that govern offline life.

Moreover, Balkin's call to embed free speech values into policy and the design of digital infrastructures resonates with Celeste's description of 'normative counteractions' that adapt constitutional logic to constrain both state and corporate power in the digital realm.²²⁰ Backed by a growing body of influential scholarship and advocacy,²²¹ digital constitutionalism offers an additional normative foundation for extending fundamental rights protection beyond the traditional state centric protections, which tightly relates to this study's declared aims.

Without intending to exhaust all relevant scholarship and all the possible parallels they presents, the perspectives brought in this chapter sharpen and strengthen Balkin's

²¹⁸ Edoardo Celeste, *Digital Constitutionalism the Role of Internet Bills of Rights*. (Routledge 2023) 79–80 <<https://www.taylorfrancis.com/books/9781003256908>> accessed 07 July 2025.

²¹⁹ Ibid, 84.

²²⁰ Ibid.

²²¹ Amongst them, developing how constitutionalism—in particular rule of law values—can be used to evaluate legitimacy of governance by platforms: Nicolas P. Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms.' (2018) 4(3) *Social Media + Society* 1. <<https://doi.org/10.1177/2056305118787812>> accessed 15 August 2025.

framework, presenting additional normative justification for an actualized vision of free speech and free speech regulation, capable of confronting the challenges of the digital age. A shared understanding that our possibilities of expression are simultaneously enabled and constrained by the material infrastructures they travel on, and that neoliberal platform capitalism is antithetical to democratic discourse, underscores the need for structural interventions that move beyond superficial procedural oversight to secure expressive freedom as a living, democratic practice. The intersection of legal scholarship with works in the field of communication and computational sciences not only affirms the urgency of structural interventions but also outlines its contours, underscoring the need to address the economic logics that drive distortions in the digital public sphere.

Chapter 3 – Examining Concrete Regulatory Measures for Structural Reform

Complemented by an interdisciplinary body of scholarship, Balkin’s framework for infrastructure free speech protection, is capable of informing a new regulatory paradigm for algorithmic governance—one that directly acknowledges the constraints social media places on free expression and democratic participation and seeks to counter them deliberately. Such a paradigm would be resolute in advancing non-discriminatory popular participation, ensuring equitable access to audiences, and fostering democratic control over the material infrastructures that enable participation and cultural production. It would also raise above industry myths and standards designed to constrain regulatory ambition and address the amplification of social pathologies as direct consequences of platform business model and power asymmetry.

Drawing primarily from Hartzog’s “dead-ends, design, duties, and defaults”²²² model for AI governance, this final chapter will explore what this paradigm shift could look like. Unapologetically attuned to the realities of platform capitalism, Hartzog advances a model that insists on enforceable boundaries targeting its economic, architectural, and ideological foundations—confronting platform dominance with more than superficial procedural rules. Although the measures Hartzog proposes were originally conceived to address data governance concerns, through dialogue with Balkin’s work and other

²²² Hartzog (n 98).

intersecting scholarship on free speech, can be adapted to counter the distortive pressures platforms exert over public discourse and free speech, a challenge this final chapter attempts to undertake.

Dead-ends

The first pillar in Hartzog’s model—dead-ends, or outright prohibitions—serves the critical function of circuit breaking the technology normalization cycle outlined in previous chapters, providing an indispensable backstop against eroding social norms. Before asking how algorithmic systems that determine the reach and visibility of speech on the digital public sphere should be regulated, policymakers must first confront the fundamental question of whether they should exist at all.²²³ These systems were created, not for our benefit, but to allow platforms to commodify speech and user attention for profit. Although users do benefit from accessing content that is personalized and ranked to their liking, regulators should not exempt themselves from asking whether we might be collectively better off without these systems.

The AI Act’s striking omission of content distribution systems—mirrored by their abrupt removal from the Brazilian AI bill—illustrates the political challenges of imposing real constraints on platform power. These examples also underscore the urgency of empowering policymakers to exercise the full extent of their mandate. This includes banning systems and policies that fundamentally conflict with our deepest constitutional commitments when less restrictive alternatives prove ineffective or impossible. The first pillar in Hartzog’s framework reminds us of a critical truth: it is AI deployment that must conform to our existing social and normative frameworks—not the other way around. Reversing this logic has trapped policymakers in a paradigm where regulation merely legitimizes prevailing technological practices, rather than asserting the primacy of democratic values over economic gain. If we do not manage to find a way out, we risk abandoning our last line of defence against platform dominance.

Does this mean that regulators should simply ‘nip the problem in the bud’ and prohibit the use of predictive AI for content distribution entirely? Or should they prefer a more nuanced approach to the construction of dead-ends, for example, restricting the use of

²²³ Ibid 641.

predictive AI within a user's own network of followers to prevent the amplification of content from non-followed accounts? Given the inherent limitations of this study, we cannot intend to provide a definitive answer on these questions. Rather, our aim is to reinforce the validity and importance of dead-ends as a regulatory tool and to empower regulators to examine and pursue it when necessary. Although they should never be a first resort, dead-ends are not only possible but have historically enabled significant advances in other fields like consumer and environmental protection—all without suppressing liberty and economy—as previously outlined.

Just as laws worldwide prohibit hazardous chemicals, like lead-based paint, in consumer products, or restrict the disposal of toxic waste into rivers to protect public health, regulators should likewise take seriously the prospect of banning algorithmic systems that undermine fundamental free speech rights and democratic participation. Especially when such consequential effects produce little more than profit and platform power. Such exercise must, of course, observe all relevant legal and constitutional constraints, including ones that might emerge through the pertinent mechanisms of constitutional balancing (e.g., proportionality principle).

Design rules

As outlined in earlier chapters, the neoliberal assumptions that innovation will invariably bring collective good and that AI tools are essentially amoral have profoundly shaped digital regulation. This has not only fostered an overreliance on procedural safeguards but has also stigmatized design rules as overly invasive or economically stifling, making it politically taboo to even discuss restrictions on algorithmic design or the underlying incentive structures that shape it. Yet, algorithms are not amoral tools; they are products of human choices and institutional processes; and they affect our social reality, making certain outcomes more or less likely.²²⁴

Driven for profit and data extraction, the predictive systems that organize content distribution on social media will likely do more collective harm than good—unless regulated proactively, intentionally and effectively. Recognizing this danger, Hartzog emerges as an influential proponent of binding design rules to constrain platform

²²⁴ Hartzog (n 98) 636.

arbitrariness, anchoring a key pillar of his AI governance model on this idea. He contends that by sending signals and making tasks either easier or harder, design choices affect the creation and distribution of power.²²⁵ Redistributing platform power therefore requires regulators to actively engage in algorithmic design, mandating outcomes and setting constraints that align technology with democratic values.

Hartzog's vision, initially conceived to promote data governance, directly aligns with Balkin's conclusion that preserving free speech requires actively embedding technological design with free speech values. Yet, the imperative to develop such rules has, thus far, been largely neglected by regulators. In the EU, the DSA and the DMA's fundamentally procedural regimes make no attempt at restricting or orienting the algorithmic design of 'recommender systems', leaving platform discretion unchecked. The AI Act imposes only a single and subtle restriction on AI design altogether. Article 14(1) demands high-risk applications be designed to allow effective human oversight. However, as previously discussed, this requirement does not explicitly apply to content distribution, seeing as this application has been entirely omitted from the Act's risk classification. Furthermore, although technically a design rule, the regulatory focus of allowing human oversight remains confined to procedural transparency rather than extending to mandating outcomes.

Rather than focusing on surface level adjustments, Hartzog suggests regulators might develop more significant design rules through Cohen's 'infrastructure thinking'²²⁶ model. As previously outlined, this concept allows regulators to assess the underlying arrangements that enable and shape large-scale exchanges to identify and address the root causes of platform harm beyond surface-level symptoms. In the context of content distribution, this expansive analytical lens might allow regulators to challenge features that are deeply entrenched in platforms' business model, either establishing democratic goals—like youth protection, or the amplification of historically and economically marginalized voices—as overriding factors in content distribution, or inspiring creative solutions to circumvent and compensate platform defaults.

²²⁵ Ibid.

²²⁶ Ibid 638.

As previously acknowledged, a key perspective integrating the issues of algorithmic harm, design regulation and free speech protection emerges from Langvardt's work on the online public sphere. In it, he outlines a multilevel strategy that entails on one hand, targeting the mechanisms responsible for propagating harm internally. This might include restructuring data monetization practices, eliminating features that amplify extreme content and promoting context-sensitive interventions (e.g., strengthen the increasingly precarious position of journalists and news media outlets on social media).²²⁷ On the other hand, Langvardt also emphasizes external measures to decentralize platform control. He points, for instance, to promoting direct creator-audience connections through subscription models, public funding, or decentralized infrastructure that would act as a stabilizing force to platform dominance.²²⁸

In this context, Langvardt examines a particularly innovative solution: promoting algorithmic portability as a means of circumventing platforms' engagement-optimizing designs entirely.²²⁹ By allowing users to import and apply their own preferred ranking systems into their social media profiles, these mechanisms might obviate the need for extensive restrictions on the platforms themselves. And by enabling users to choose, or even develop, algorithms that reflect their own civic values, rather than platforms' monolithic profit-driven defaults—this model could potentially inaugurate a new era of user agency and pluralism in digital discourse with minimal intervention.

There are indeed multiple paths through which to approach design regulation beyond surface level interface adjustments in the context of content distribution. Once we abandon the illusion that minor tweaks can meaningfully reform extractive systems and the fallacy that external oversight should never challenge platforms underlying business model, new and transformative possibilities emerge. Each of the possibilities examined above would theoretically help reduce platform control over cultural flows and virality—either by introducing public interest goals directly into algorithmic designs, or by promoting external solutions to compensate platform power. As Langvardt has demonstrated, even external measures that do not directly affect platforms'

²²⁷ Ibid.

²²⁸ Ibid.

²²⁹ Langvardt (n 3) 30.

operations—such as promoting algorithmic portability—can meaningfully enhance user agency and pluralism, thereby strengthening free speech and democratic participation.

Once again, we do not intend to provide a definitive answer on how regulators should proceed. Rather, our aim is to shed light on the possibilities that unveil once we re-empower regulators to pursue their mandate to its full constitutional extent. Genuine progress requires decentralizing platform control over digital discourse and implementing structural reforms that redistribute power. Design mandates are indispensable to this effort, serving as the means through which free speech values—such as interactivity, equitable access, democratic oversight, and participatory agency—can be embedded directly into technological infrastructures. Such measures, however, must always be implemented in line with legal and constitutional constraints, particularly the principle of proportionality.

Duties

Hartzog's model for AI governance also comprises the creation of binding, non-negotiable duties of loyalty and care towards data governance. In developing his vision, Hartzog proposes a two-step approach. First, the articulation of a primary, general duty of loyalty, and second, the development of subsidiary duties that are more specific and context sensitive. The intended result is a framework combining flexible, open-ended obligations with precise, enforceable rules tailored to targeting the most opportunistic and harmful instances of platform self-dealing.²³⁰

In the context of content distribution, secondary, context-specific duties could be formulated based on the various design alternatives previously discussed. This can be accomplished by simply shifting the focus from aspirational design principles to concrete, binding legal mandates. For instance, instead of establishing design restrictions to favour historically marginalized voices, and to avoid the circulation of divisive content, regulator might articulate a secondary liability regime to enforce a predetermined standard for democratic participation that prohibits the systematic suppression of less profitable speech, or the algorithmic amplification of divisive content for engagement purposes. This approach would help align platform governance to the

²³⁰ Hartzog (n 98) 634.

crucial ethical standards that traditionally guide professional journalism and to liability frameworks that have long governed mass media advertising on other medias.

A similar proposal, based on a framework of clear professional duties, is also present in Balkin's reconceptualization of social media platforms as 'information fiduciaries'—entities bound by strict duties of loyalty, care and confidentiality, akin to the fiduciary obligations of doctors or lawyers to their clients.²³¹ In justifying this framework, Balkin explains that "new obligations arise from people's increasing dependence on and vulnerability to digital services that collect data about them but whose operations are not transparent to them."²³² While Balkin's approach to ground fiduciary duties in trust and legitimate user expectations is compelling, his model also presents significant practical limitations. Notably, it argues that fiduciary obligations could not be state imposed, rather, they should emerge voluntarily from the industry itself.²³³

The assumption that platforms will one day willingly prioritize user welfare over their core profit engine is fundamentally flawed. Unlike traditional fiduciaries—whose professions are structurally constrained by ethical codes, licensure requirements, and institutional oversight—platforms operate within a system that actively rewards data extraction, algorithmic manipulation, and engagement maximization. Although Balkin's information fiduciary model offers valuable insights, his reluctance to acknowledge the economic reality of platform capitalism introduces a critical vulnerability that regulators must overcome. True reform requires clear, enforceable standards of conduct that directly address the structural incentives of platform economics and acknowledge the documented historical failure of self-regulatory and co-regulatory schemes.

The articulation of duties that bind platform practices would not contradict but rather complement efforts aimed at restricting and guiding algorithmic design. While both approaches theoretically pursue similar ends—curbing harmful platform practices and reinforcing democratic values—they operate through distinct regulatory levers. The articulation of binding duties offers a normative and enforceable foundation, whereas design mandates intervene more directly in the technological architectures that mediate

²³¹ Balkin (n 19) 2048.

²³² Ibid.

²³³ Ibid 2043.

speech. Ideally, these strategies should work in tandem. However, further academic inquiry and regulatory experimentation are needed to determine how best to calibrate and integrate these approaches to ensure they are both effective and adaptable across different governance contexts.

Defaults

Finally, in a decisive critique of the reactive self-assessment regimes that underpin the current paradigm of digital regulation, Hartzog calls for a more assertive and proactive mechanism for assessing algorithmic harm. To this end, he proposes drawing inspiration from the European Union’s General Data Protection Regulation (GDPR)—particularly the foundational default that all data processing is presumed unlawful unless affirmatively justified by a legal basis.²³⁴

Applied to content distribution, this approach might require platforms to proactively demonstrate that their systems are neither discriminatory, manipulative, unfair, nor inaccurate, and that they comply fully with fundamental rights protections, under the penalty of being deemed unlawful. By reversing the existing presumption of legality, this model would shift the burden of proof squarely onto platforms, thereby confronting the DSA’s problematic reliance on *ex-post* reporting and self-assessment for risk management. This vision is also promoted by Sandra Wachter. In her critical assessment of the AI Act, she challenges the notion of neutral data, emphasizing that the datasets used to train AI systems inevitably reflect longstanding biases embedded in sectors such as employment, finance, education, criminal justice, and healthcare. Accordingly, Wachter calls for a reversal of the burden of proof, requiring providers to demonstrate that their systems are in fact unbiased.²³⁵

The development of such defaults—embedding legal accountability into the design and deployment pipeline—aligns with Balkin’s vision that safeguarding free speech in the digital age requires injecting interactivity, equitable access, democratic oversight, and participatory agency directly into the technical architectures that shape online discourse. By integrating these normative commitments at an infrastructural level,

²³⁴ Hartzog (n 98).

²³⁵ Wachter (n 86) 689.

regulation can transcend reactive models and move toward a more anticipatory, structural approach to algorithmic speech governance—one that is better suited to meet the demands of a truly democratic society.

Preventive, *ex-ante* controls have long been a cornerstone in environmental protection, where they have proven effective both in averting systemic risks and in facilitating more efficient enforcement. The application of a similar approach to digital regulation could transform how we address the harms of algorithmic governance, allowing a shift from risk mitigation to harm prevention. Once we acknowledge the failure of current *ex post* risk assessment mechanisms—which often lack clear benchmarks and are easily manipulated by companies to simulate compliance while concealing harmful practices—stronger, more proactive measures are not only welcome but essential for effective oversight.

Concluding Remarks

This study opens by examining a key paradox of the digital age: how platforms once praised for enabling democratic expression have become powerful and unaccountable gatekeepers of participation. By combining surveillance, engagement-optimization and predictive targeting, social media has engineered algorithms that systematically privilege provocative, polarizing, and captivating content—undermining the participatory values of free speech and democratic deliberation in the pursuit of profit. From rampant disinformation and hate speech to polarization and general social fragmentation, these systems have been frequently linked to the pathologies that now define the digital age.

This study argues that the consolidation of these systems has fundamentally shifted the battleground of free speech. The central challenge to free expression is no longer solely overt censorship, but the pervasive algorithmic manipulation of the conditions that allow expression to matter in the first place—namely, our ability to reach an audience, participate in cultural production and meaningfully contribute to the narratives that shape society and our collective and individual identities. The central normative challenge we address is how to counter this control and protect public discourse from

the commodification pressures that govern content circulation and increasingly shape social and political agendas in the real world.

Our analysis demonstrates that the prevailing regulatory and conceptual frameworks remain woefully inadequate in tackling this challenge. In the EU, the AI Act's striking omission of content distribution systems, the DSA's reliance on performative transparency and self-assessment regimes, and the DMA's narrow focus on the economic consequences of ranking discrimination collectively amount to regulatory failure. Shaped by constitutional constraints, industry lobbying, and theoretical misconceptions that fetishize innovation and algorithmic neutrality, these frameworks privilege procedural compliance and market efficiency over democratic participation and free speech—leaving the digital public sphere exposed to systemic commodification and manipulation.

To break this cycle, this paper proposes a dual path forward. First, drawing from the works of Jack M. Balkin, Julie E. Cohen, Kayle Langvardt and Kai Riemer and Sandra Peter, it offers a theoretical model for reconceptualizing free speech within platform and algorithmic governance, supporting the development of stronger protections that recognize the unique custodial role of social media in enabling participation and cultural production. Second, inspired by Woodrow Hartzog's framework for AI governance, it presents an analysis of concrete, proactive, rights-based measures capable of realizing this vision.

The journey toward a democratically aligned digital public sphere is undoubtedly complex, fraught with political, technical, and constitutional challenges. This study does not presume to offer a single, definitive solution but rather a coherent roadmap for navigating the challenges of insulating public discourse from the commodification pressures of platform business model. By diagnosing the failures of current frameworks and synthesizing insights from law, political science, and communication studies, we have provided a normative vision to guide reform and a toolkit to achieve it.

More than anything, the path forward requires rejecting the fatalistic acceptance of the commodification of speech as an inevitable cost of progress. We must affirm that our technological future is not predetermined; it is a collective choice. By empowering

regulators to act with theoretical clarity and political courage, we can choose a future where technology serves democracy, and where the digital public sphere finally fulfils its promise as a space for genuine connection, equitable participation, and collective self-determination.

Bibliography

Cases

Handyside v United Kingdom App no 5493/72 (ECHR, 1 December 1976)

EU Legislation & Official Documents

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, 2014/53/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 1/1

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) [2022] OJ L 277/1

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) [2022] OJ L 265/1

European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act)' (Explanatory Memorandum, COM(2020) 842 final, 15 December 2020) <<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0842>> accessed 15 August 2025

Other Legislation & Official Documents

Comissão Temporária Sobre Inteligência Artificial no Brasil, *Parecer 1/2024* (5 December 2024) <https://legis.senado.leg.br/sdleg-getter/documento?dm=9858756&ts=1738768138734&rendition_principal=S&disposition=inline> accessed 15 August 2025

Comissão Temporária Sobre Inteligência Artificial no Brasil, *Parecer 208/2024* (10 December 2024) <<https://legis.senado.leg.br/sdleg-getter/documento?dm=9884955&ts=1742240908365&disposition=inline>> accessed 15 August 2025

Comissão Temporária Sobre Inteligência Artificial no Brasil, *Parecer s/n* (7 June 2024) <https://legis.senado.leg.br/sdleg-getter/documento?dm=9630164&ts=1742240894034&rendition_principal=S&disposition=inline> accessed 15 August 2025

List of References

AlgorithmWatch, 'Busted Internet Myth: Algorithms Are Always Neutral' (25 November 2019) <<https://algorithmwatch.org/en/busted-internet-myth-algorithms-are-always-neutral/>> accessed 15 August 2025

Amnesty International, *Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations – New Report* (29 September 2022) <https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/> accessed 14 August 2025

Armijo, E., 'Speech Regulation by Algorithm' (2021) 30 *William & Mary Bill of Rights Journal* 245 <https://scholarship.law.wm.edu/wmborj/vol30/iss2/3> accessed 15 March 2025

Balkin, J.M., 'Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society' (2004) 79 *New York University Law Review* 1 <https://jackbalkin.yale.edu/sites/default/files/files/digitalspeechanddemocraticculture1.pdf> accessed 11 August 2025

Balkin, J.M., 'Free Speech is a Triangle' (2018) Yale Law School Public Law Research Paper No 640 <https://ssrn.com/abstract=3186205> accessed 15 August 2025

Balkin, J.M., 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 *UC Davis Law Review* 1148 <https://openyls.law.yale.edu/server/api/core/bitstreams/2a57c714-e278-449a-8ba0-91c6d56c34e2/content> accessed 20 August 2025

Benkler, Y. and others, 'Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics' (2018) *Oxford University Press*

<https://library.oapen.org/viewer/web/viewer.html?file=/bitstream/handle/20.500.12657/28351/9780190923624.pdf?sequence=1&isAllowed=y> accessed 22 July 2025

Botti, F. and others 'The #Metoo Social Media Effect and its Potentials for Social Change in Europe' (Foundation for European Progressive Studies, 2019) <https://feps-europe.eu/wp-content/uploads/2019/10/The-Metoo-Social-Media-Effect-and-its-Potentials-for-Change-in-Europe.pdf> accessed 12 August 2025

Brownstein, M., 'Social Media Platforms Generate Billions in Annual Ad Revenue from U.S. Youth' (Harvard T.H. Chan School of Public Health, 27 December 2023) <https://hsph.harvard.edu/news/social-media-platforms-generate-billions-in-annual-ad-revenue-from-u-s-youth/> accessed 14 August 2025

Buiten, M.C., 'The Digital Services Act: From Intermediary Liability to Platform Regulation' (2021) 12 *JIPITEC* 361 <https://www.ssrn.com/abstract=3876328> accessed 14 June 2025

Calo, R., 'Digital Market Manipulation' (2014) 82 *George Washington Law Review* 995 <https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=1024&context=faculty-articles> accessed 14 August 2025

Carrieri, V. and others, 'Vaccine Hesitancy and (Fake) News: Quasi-Experimental Evidence from Italy' (2019) *Health Economics Letter* <https://doi.org/10.1002/hecl.3937> accessed 22 July 2025

Celeste, E., *Digital Constitutionalism the Role of Internet Bills of Rights* (Routledge 2023) <https://www.taylorfrancis.com/books/9781003256908> accessed 07 July 2025

Chang, H. and others, '#JusticeforGeorgeFloyd: How Instagram Facilitated the 2020 Black Lives Matter Protests' (2022) *PLOS ONE* <https://doi.org/10.1371/journal.pone.0277864> accessed 19 August 2025

Cho, J. and others, 'Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization.' (2020) 64(2) *Journal of Broadcasting & Electronic Media*, 150 <https://doi.org/10.1080/08838151.2020.1757365> accessed 22 July 2025

Codagnone, C. and others, 'Leading the Charge on Digital Regulation: The More, the Better, or Policy Bubble?' (2023) 2 *DISO* 1 <https://doi.org/10.1007/s44206-023-00033-7> accessed 15 August 2025

Cohen, J.E., *Between Truth and Power the Legal Constructions of Informational* (Oxford University Press 2019) <https://doi.org/10.1093/oso/9780190246693.001.0001> accessed 13 August 2025

Cohen, J.E., 'Infrastructuring the Digital Public Sphere' (2023) 25 *Yale JL & Tech* 1 https://yjolt.org/sites/default/files/cohen_julie_infrastructuring_the_digital_public_sp_here.01.pdf accessed 15 August 2025

D'Cruze, D., 'Elon Musk Confirms That X Deprioritises Posts With Links To Discourage "Lazy Linking"' *Business Today* (26 November 2024) <https://www.businesstoday.in/technology/news/story/elon-musk-confirms-that-x-deprioritises-posts-with-links-to-discourage-lazy-linking-455115-2024-11-26> accessed 15 August 2025

Darcy, O., 'Publishers See Dramatic Drop in Facebook Referral Traffic as the Social Platform Signals Exit from News Business' *CNN* (17 August 2023) <https://edition.cnn.com/2023/08/17/media/facebook-referral-traffic-reliable-sources> accessed 15 August 2025

Silva, V.D.N. and others, 'Are Algorithms Affecting the Democracy in Brazil?' (International Symposium on Ethical Algorithms, December 2018) https://www.researchgate.net/profile/Regina-Alves-Da-Silva/publication/332798661_Are_Algorithms_Affecting_the_Democracy_in_Brazil/links/5cca355d4585156cd7c1ac22/Are-Algorithms-Affecting-the-Democracy-in-Brazil.pdf accessed 22 July 2025

Dubai School of Government, *Arab Social Media Report* (Vol 1, No 2, May 2011) https://journalistsresource.org/wp-content/uploads/2011/08/DSG_Arab_Social_Media_Report_No_2.pdf accessed 15 August 2025

Ebers, M., 'Truly Risk-based Regulation of Artificial Intelligence How to Implement the EU's AI Act' (2024) 15 *European Journal of Risk Regulation* 1 <https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/truly-riskbased-regulation-of-artificial-intelligence-how-to-implement-the-eus-ai-act/E526C1D0D7368F9691082220609D60F4> accessed 14 May 2025

Edelson, L. and others, 'Far-Right News Sources on Facebook More Engaging' (Medium, 2021) <https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging04a01efae90> accessed 21 February 2025

European Commission, 'The Digital Markets Act is the EU's Law to Make the Markets in the Digital Sector Fairer and More Contestable' (14 February 2024) https://digital-markets-act.ec.europa.eu/index_en accessed 15 August 2025

European Commission, 'Commission Seeks Input to Clarify Rules for General-Purpose AI Models' (Press Release, 22 April 2024) <https://digital-strategy.ec.europa.eu/en/news/commission-seeks-input-clarify-rules-general-purpose-ai-models> accessed 14 May 2025

Facebook Help Centre (updated 11 June 2024) <https://www.facebook.com/help/1257205004624246/> accessed 14 August 2025.

Gal, M. S., 'Algorithmic Challenges to Autonomous Choice' (2018) 25 *Michigan Technology Law Review* 59 <https://repository.law.umich.edu/mtlr/vol25/iss1/3> accessed 15 August 2025

Gillespie, T., *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018) <https://doi.org/10.12987/9780300235029> accessed 18 August 2025

Goel, V., 'Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry' (*The New York Times*, 29 June 2014) <https://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html> accessed 14 June 2025

Grozdanovski, L. and Cooman, J., 'Forget the Facts, Aim for the Rights! On the Obsolescence of Empirical Knowledge in Defining the Risk/Rights-Based Approach to AI Regulation in the European Union.' (2023) 49 *Rutgers Computer & Technology Law Journal* 208

<https://orbi.uliege.be/bitstream/2268/307355/1/Grozdanovski%20and%20De%20Cooman%2C%20Forget%20the%20Facts%2C%20Aim%20for%20the%20Rights%20-%2049.2%20RutgersComputerTechLJ.pdf> accessed 15 August 2025

Hagey, K. and Horwitz, J., 'Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead.' *The Wall Street Journal* (15 of September 2021) https://www.wsj.com/tech/facebook-algorithm-change-zuckerberg-11631654215?mod=article_inline accessed 14 August 2025

Hartzog, W., 'Two AI Truths and a Lie' (2023) 26 *Yale Journal of Law & Technology* 596 https://vjolt.org/sites/default/files/hartzog_26yalejltech595.pdf accessed 14 August 2025

Hill, K., 'Facebook Manipulated 689,003 Users' Emotions For Science' (*Forbes*, 28 June 2014) <https://www.forbes.com/sites/kashmirhill/2014/06/28/facebook-manipulated-689003-users-emotions-for-science/> accessed 14 July 2025

Hobbelen, H. and Wouters, D., 'First-ever DMA Non-Compliance Fines -- Another Step in Big Tech Scrutiny' (*Competition Law Insights*, 20 May 2024) <https://competitionlawinsights.twobirds.com/post/102kaib/first-ever-dma-non-compliance-fines-another-step-in-big-tech-scrutiny> accessed 15 August 2025

Horwitz, J., 'Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.' *The Wall Street Journal* (13 September 2021) https://www.wsj.com/tech/facebook-files-xcheck-zuckerberg-elite-rules-11631541353?mod=article_inline accessed 14 August 2025

Huszár, F. and others 'Algorithmic Amplification of Politics on Twitter' (2022) 119(1) *Proceedings of the National Academy of Sciences* e2025334119 <https://doi.org/10.1073/pnas.2025334119> accessed 15 August 2025

Jones, M.G., 'Tech Companies Spend More Than €100 Million a Year on EU Digital Lobbying' (*EuroNews*, 11 of September 2023) https://www.euronews.com/my-europe/2023/09/11/tech-companies-spend-more-than-100-million-a-year-on-eu-digital-lobbying?utm_source=chatgpt.com accessed 15 August 2025

Kapczynski, A., 'The Law of Informational Capitalism' (2020) 129 *Yale Law Journal* 1460 https://www.yalelawjournal.org/pdf/KapczynskiBookReview_iqh4qxtw.pdf accessed 15 August 2025

Langvardt, K., 'A New Deal for the Online Public Sphere' (2019) 26 *George Mason Law Review* 1 <https://ssrn.com/abstract=3149513> accessed 11 August 2025

León, L.P., 'Big Tech é excluída do rol de Inteligência Artificial de alto risco' (*Agência Brasil*, 2024) <https://agenciabrasil.ebc.com.br/politica/noticia/2024-12/big-tech-e-excluida-do-rol-de-inteligencia-artificial-de-alto-risco> accessed 15 August 2025

Madiega, T., *Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act* (European Parliamentary Research Service, May 2020) PE 649.404 [https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA\(2020\)649404_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA(2020)649404_EN.pdf) accessed 15 June 2024

'Mark Zuckerberg asks governments to help control internet content' (*BBC News*, 30 March 2019) <https://www.bbc.com/news/world-us-canada-47762091> accessed 14 July 2025.

Matias, N. and Wright, L., 'Impact Assessment of Human-Algorithm Feedback Loops' (1 of March 2022) *Just Tech* <https://doi.org/10.35650/JT.3028.d.2022> accessed 14 August 2025

Meta, *Systemic Risk Assessment and Mitigation Report for Facebook* (Pursuant to Regulation (EU) 2022/2065, August 2024) <https://transparency.meta.com/reports/regulatory-transparency-reports/> accessed 20 May 2025

Meta, *Meta Consumer Profiling Techniques: Digital Markets Act Non-Confidential Public Overview* (6 March 2025) <https://transparency.meta.com/sr/meta-consumer-profiling-2025> accessed 15 August 2025.

Meta Newsroom, 'News Feed FYI: Taking Into Account Time Spent on Stories' (Facebook, 12 June 2015) <https://about.fb.com/news/2015/06/news-feed-fyi-taking-into-account-time-spent-on-stories/> accessed 11 August 2025

Meta Newsroom, 'Bringing People Closer Together' (11 January 2018) <https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/> accessed 12 May 2025

Meta Transparency Centre, 'Explaining Ranking on Facebook' (Transparency Centre, 31 December 2023, last updated 31 December 2023) [https://transparency.meta.com/features/explaining-ranking/#facebook ai systems](https://transparency.meta.com/features/explaining-ranking/#facebook_ai_systems) accessed 16 May 2025

Meta Transparency Centre, 'Our Approach to Facebook Feed Ranking' (Meta Transparency Centre, updated 11 June 2025) <https://transparency.meta.com/pt-br/features/ranking-and-content/> accessed 27 August 2025.

Meßmer, A. and Degeling, M., 'Auditing Recommender Systems -- Putting the DSA into Practice With a Risk-Scenario-Based Approach' (Stiftung Neue Verantwortung, February 2023) <https://doi.org/10.48550/arXiv.2302.04556> accessed 14 May 2025

Minkin, A., 'As Washington Debates Major Tech and AI Policy Changes, Big Tech's Lobbying is Relentless.' (*Issue One*, 22 July 2025) https://issueone.org/articles/as-washington-debates-major-tech-and-ai-policy-changes-big-techs-lobbying-is-relentless/?utm_source=chatgpt.com accessed 15 August 2025

Narayanan, A., 'Understanding Social Media Recommendation Algorithms: Towards a Better Informed Debate on the Effects of Social Media' (Knight First Amendment Institute at Columbia, 2023) <http://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms> accessed 27 May 2025

Papacharissi, Z., 'The Virtual Sphere the Internet as a Public Sphere.' (2002) 4 *New Media & Society* 9 <https://doi.org/10.1177/14614440222226244> accessed 15 August 2025

Guniganti, P., 'Alexander Meiklejohn: Teacher and Citizen' (2005) Ideas of the First Amendment (Thomson/West 2006) <<https://ssrn.com/abstract=2104006>> accessed 23 May 2024

Parida, S., 'Evolution of Social Media Algorithms: The Invisible Hand Guiding Our Online Experience' *Reso* (6 October 2024) <https://resoinsights.com/insight/evolution-of-social-media-algorithms/> accessed 15 August 2025

Pathways: How Digital Design Puts Children at Risk (5Rights Foundation, July 2021) <https://5rightsfoundation.com/wp-content/uploads/2021/09/Pathways-how-digital-design-puts-children-at-risk.pdf> accessed 14 August 2025

Protecting Youth Mental Health: The U.S. Surgeon General's Advisory (US Department of Health and Human Services, 2021) <https://www.hhs.gov/sites/default/files/surgeon-general-youth-mental-health-advisory.pdf> accessed 14 August 2025

Riemer, K. and Peter, S., 'Algorithmic Audiencing: Why we Need to Rethink Free Speech on Social Media' (2021) *Journal of Information Technology* 409 <http://journals.sagepub.com/doi/10.1177/02683962211013358> accessed 13 August 2025

Siapera, E. and Viejo-Otero, P., 'Governing Hate: Facebook and Digital Racism' (2021) 22(2) *Television & New Media* 112 <https://doi.org/10.1177/1527476420982232> accessed 22 July 2025

Soler, P., 'Big Tech, Banking, Energy: Who Are the Biggest Spenders on EU Lobbying?' (*EuroNews*, 24 of February 2025) https://www.euronews.com/my-europe/2025/02/24/big-tech-banking-energy-who-are-the-biggest-spenders-on-eu-lobbying?utm_source=chatgpt.com accessed 14 August 2025.

Stecklow, S., 'Why Facebook is Losing the War on Hate Speech In Myanmar' *Reuters Investigates* (15 August 2018) <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> accessed 22 July 2025

Suzor, N.P., 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4(3) *Social Media + Society* 1. <https://doi.org/10.1177/2056305118787812> accessed 15 August 2025.

‘The Constitution: Supreme Court Case Library: *Abrams v United States*’ (National Constitution Center) <https://constitutioncenter.org/the-constitution/supreme-court-case-library/abrams-v-united-states> accessed 21 February 2025

Turvy, A., ‘Comparing Tiktok and Instagram’s Sociotechnical Environments for Cultural Production’ (2025) 2 *Platforms & Society* 1 <https://doi.org/10.1177/29768624251359796> accessed 20 August 2025

Protecting Youth Mental Health: The U.S. Surgeon General's Advisory (US Department of Health and Human Services, 2021) <https://www.hhs.gov/sites/default/files/surgeon-general-youth-mental-health-advisory.pdf> accessed 14 August 2025

TikTok Inc., *Community Guidelines Enforcement Report: First Half 2025* (26 June 2025) <https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2025-1> accessed 14 May 2025

Vosoughi, S., Roy, D. and Aral, S., ‘The Spread of True and False News Online’ (2018) 359 (6380) *Science* 1146 <https://www.science.org/doi/10.1126/science.aap9559> accessed 12 August 2025

Wachter, S., ‘The Limitations of the EU’s Artificial Intelligence Act and the AI Liability Directive’ (2024) 26 *Yale Journal of Law & Technology* 674 https://yjolt.org/sites/default/files/wachter_26yalejltech671.pdf accessed 14 May 2025

Wei, K. and others, ‘How Do AI Companies “Fine-Tune” Policy? Examining Regulatory Capture in AI Governance’ (*Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 16 October 2024) <https://doi.org/10.48550/arXiv.2410.13042> accessed 15 August 2025.

Wells, G. and others, ‘Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show.’ *The Wall Street Journal* (13 September 2021) https://www.wsj.com/tech/personal-tech/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?mod=hp_lead_pos7&article_inline accessed 14 August 2025.

Wu T., 'Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-ordering Systems' (2019) 119 *Columbia Law Review* 2001
<https://www.jstor.org/stable/10.2307/26810857> accessed 15 March 2025

X Corp., *Global Transparency Report: January - June 2024* (February 2025)
<https://transparency.x.com/en/reports/global-reports/2025-transparency-report>
accessed 14 May 2025