SEEING WHAT OTHERS ARE SEEING

Studies in the regulation of transparency for social media recommender systems

Paddy Leerssen

SEEING WHAT OTHERS ARE SEEING

Studies in the regulation of transparency for social media recommender systems

by

Paddy Leerssen

Cover Photo: Machinery of the printing house Kaleva - Finnish Heritage Agency, Finland - CC BY. https://www.europeana.eu/item/2021009/_FB1E511127B67BAC69960049D0410CE2

Seeing What Others Are Seeing

Studies in the regulation of transparency for social media recommender systems

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 12 april 2023, te 14.00 uur

door Patrick James Leerssen
geboren te Utrecht

Promotiecommissie

Promotores:	prof. dr. N. Helberger	Universiteit van Amsterdam
	prof. dr. C.H. de Vreese	Universiteit van Amsterdam
Copromotores:	prof. dr. T. McGonagle	Universiteit van Amsterdam
Overige leden:	prof. dr. M.R.F. Senftleben	Universiteit van Amsterdam
	dr. J.V.J. van Hoboken	Universiteit van Amsterdam
	prof. dr. T. Poell	Universiteit van Amsterdam
	prof. dr. J.F.T.M. van Dijck	Universiteit Utrecht
	prof. dr. W. Schulz	Universität Hamburg
	dr. M.E. Kaminski	University of Colorado

Faculteit der Rechtsgeleerdheid

Table of contents

Acknowledgements **10**

CHAPTER 1 | Introduction **13**
1. Introduction **15**
2. Background: The problem of transparency in social media recommender systems **18**
2.1 Governance *of* and *by* social media platforms 18
2.2 Recommender systems as points of control in social media governance ... 23
2.3 Of what and for whom? Transparency in recommender systems 29
3. Research question and outline **34**
4. Methods and Format **36**
4.1 Methods 36
4.2 Format 39

CHAPTER 2 | The soap box as a black box:Regulating transparency in social media recommender systems **43**
1. Introduction **45**
2. Social media recommenders systems as opaque gatekeepers of online content **46**
2.1 What are social media recommender systems? 46
2.2 Recommendation governance: From the attention economy to attention politics 48
2.3 'Obscured obscuring': The opacity of social media recommendations 53
3. State of play: Regulating recommendation transparency in Europe **58**
3.1 User-facing disclaimers 61
3.2 Government oversight 65
3.3 Research partnerships with academia and civil society 70
4. The case for public access **74**
4.1 The pros and cons of public access 75
4.2 Designing public access 77
4.3 Regulating public access 82
5. Conclusion **84**

CHAPTER 3 | Platform ad archives: promises and pitfalls **87**
Abstract **88**
1. Introduction **89**
2. Promises: the case for ad archives **90**

2.1 Conceptual framework: what are 'ad archives'? 90
2.2 Legal framework: why are platforms building archives? 91
2.3 Normative framework: what are the policy grounds for ad archives? 96
3. Pitfalls: key challenges for ad archive architecture 99
3.1 Scoping: what ads are included in the archive? 100
3.2 Verifying: how do archives account for inauthentic behaviour? 104
3.3 Targeting: how is ad targeting documented? 107
4. Conclusion 109

CHAPTER 4 | News from the ad archive:How journalists use the Facebook Ad Library to hold online advertising accountable 113
Abstract 114
1. Introduction 115
2. Background 115
2.1 The Ad Library and its features 115
2.2 Background and rationale 117
2.3 The Ad Library as a tool for watchdog journalism 118
3. Content analysis 120
3.1 Methods 120
3.2 Pilot study 120
3.3 Content Analysis Protocol 120
3.4 Findings 123
4. Interviews 125
4.1 Method 125
4.2 Findings 127
5. Discussion 131
5.1 Impact: From publicity to accountability? 132
5.2 Critical perspectives on Ad Library journalism 134
5.3 Limitations 135
6. Conclusion 136

CHAPTER 5 | An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation 139
Abstract 140
1. Introduction 141
2. 'Shadow banning' as a function of visibility remedies 141
2.1 Definitions: what is 'shadow banning'? 142
2.2 Techniques: how do platforms shadow ban? 144
2.3 Policies: why do platforms shadow ban? 147

3. Transparency rules for content moderation in the Digital Services Act 150
3.1 The DSA's notice-and-action framework for content moderation 151
3.2 Article 14 DSA on Terms and Conditions 152
3.3 Article 17 DSA on the Statement of Reasons 155
4. Ranking due process between moderation and curation 159
4.1 Defining demotion and the problem of counterfactuals 159
4.2 Ranking transparency beyond the downrank: from moderation to curation 163
5. Conclusion 165

CHAPTER 6 | Seeing what others are seeing: Regulating social media for and with observability 169
Abstract 170
1. Introduction 171
2. Background 172
2.1 Social media regulation 172
2.2 From transparency to observability 176
2.3 Observability as a regulatory program 180
3. Regulating social media for observability 182
3.1 Observability in social media governance 182
3.2 Observability in the Digital Services Act 186
3.3 Discussion: observability of what and for whom? 189
4. Regulating social media with observability 195
4.1 Observability and regulation 195
4.2 Observability and discourse 199
5. Conclusion 201

CHAPTER 7 | Conclusions 203
1. Introduction 205
2. Transparency of what? From algorithms to sociotechnical systems 206
3. Transparency for whom? Regulating disclosure for cooperative responsibility 210
4. Conclusion 213
5. Outlook 215
5.1 Legal challenges 216
5.2 Interdisciplinary challenges 217

References **220**
Literature 220
Laws and regulations 242
Case Law 243
Soft law 243
Author contributions **246**
Summary **248**
Samenvatting **254**

Appendix I: Content Analysis Protocol **260**
Appendix II: Supplemental keyword testing **264**

ACKNOWLEDGEMENTS

In a certain sense, this entire dissertation is the result of algorithmic ranking. I still remember how, as a bachelor's student in Maastricht with a budding interest in internet governance, the top result for my Google queries would always be 'ivir.nl'. Clearly this was the place to be, I surmised, and decided to apply for IViR's master's programme, which would come to define the next decade of my life. The algorithm was right at least in this case, and for that I am grateful. However, to foreshadow some later themes in this manuscript, my real gratitude is directed not at the algorithms involved but rather at the people behind their rankings; my dear colleagues, past and present, who made IViR such a happy home for my academic upbringing.

Above all I am grateful to my supervisors. Natali, I am to this day amazed at how, amidst the countless other projects vying for your time, you were always ready to make ample room for my research. Your input was generous, inspiring, challenging when necessary, and always pushed me to reflect more deeply and avoid easy answers. Claes, I couldn't think of a better supervisor for my first forays into empirical communications research; what might otherwise have been daunting you made inviting and joyful. Tarlach, I am immensely grateful not only for your role as co-supervisor but as coach and mentor throughout my academic career.

I am also grateful to the members of my committee for their generous engagement with my work: Professors José van Dijck, Joris van Hoboken, Margot Kaminski, Thomas Poell, Martin Senftleben and Wolfgang Schulz. Special thanks are also owed to my co-author for Chapter 4, Tom Dobber, who offered invaluable guidance in the ways of quantitative empirical research.

For keeping me in good spirits throughout, I am indebted to my IViR colleagues. In particular I would like to thank Jef Ausloos, who, as office mate, co-author, and confidante, I've come to see as something of an academic big brother. I am also deeply grateful for the friendship of Martijn Sax, Jill Toh and Naomi Appelman, not only for being some of the funniest and kindest people I've had the pleasure of working with, but also for their moral clarity and conviction as young scholars and activists.

Further afield I was also lucky enough to meet generous and kind colleagues. I am especially grateful for the support, advice and friendship of Daphne Keller, Martin Husovec, Gianmarco Cristofari, Michael Veale, Rob Gorwa, Bernhard Rieder, Catalina Goanta, Heidi Tworek, Nicolo Zingales, Aleksandra Kuczerawy, Berdien van der Donk, and Evelyn Douek.

My family I cannot even begin to thank for all the countless ways they've supported me. Anna, thanks for helping me keep my sense of perspective; academia always feels a lot smaller when I think of you literally saving lives. Ann and Joep, thanks for, well, everything. I owe you the world.

Lastly there is Lotte. Of all the blessings bestowed on me in these past four years, surely the greatest is you. Your love and support throughout this process was much more than catalysis; traces of you run through every page.

It's pointless to explain why
There's nothing I can say.
My words would all be meaningless
Anyway.
It seems I can't be trusted.

- Blur, 'Explain' (1991)

CHAPTER 1
Introduction

1. Introduction

In January 2018, Facebook CEO Mark Zuckerberg decided that the platform's 2.1 billion users should see less news.[1] By adjusting their Newsfeed recommendation system, Facebook would start to prioritise content from personal connections over public pages. In defence of this so-called Meaningful Engagement policy, Facebook invoked 'feedback from our community' and assured that the change would be 'good for people's well-being'.[2] But there may well have been other motives in play. Commercially, suppressing the organic reach of public pages might encourage more spending on Facebook's advertising service.[3] Politically, suppressing news media could be a means to quell controversies around disinformation and political extremism, which had just then started to tarnish the platform's image.[4] What Facebook may not have considered, as a profit-seeking platform, are the momentous impacts on democracy and the public interest.

Zuckerberg can take such decisions because the media, along with many other societal domains, have been platformised.[5] A handful of powerful social media services now act as the 'new governors' of digital media ecosystems.[6] Their rules, inscribed in contracts and in digital technology, set the terms for public participation online.[7] A crucial point of control in this social media governance is their recommender systems: the features which

1 Zuckerberg announced this decision on through a Facebook post on his personal account. See: Mark Zuckerberg, untitled Facebook post, *Facebook*.com (12 January 2018)<https://www.facebook.com/zuck/posts/10104413015393571> accessed 26 September 2022. For user count, see: Meta, Q4 2017 Earnings Report, *Meta Investor Relations* (31 January 2018) <https://investor.fb.com/investor-events/event-details/2018/Facebook-Q4-2017-Earnings/default.aspx> accessed 26 September 2022.

2 Adam Mosseri, 'Bringing People Closer Together', Facebook Newsroom (11 January 2018). <https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/> accessed 26 September 2022.

3 Rasmus Kleis Nielsen and Sarah Ganter, *The Power of Platforms: Shaping Media and Society* (Oxford University Press 2022).

4 Yochai Benkler, Robert Faris and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press 2018). Richard Rogers and Sabine Niederer, *The politics of social media manipulation* (Amsterdam University Press 2020).

5 Anne Helmond and Fernando van der Vlist, 'Social media and platform historiography: Challenges and opportunities' (2019) 22 *TMG–Journal for Media History* 6. José van Dijck, Thomas Poell and Martijn de Waal, *The platform society: Public values in a connective world* (Oxford University Press 2018).

6 Kate Klonick, 'The New Governors: The people, rules, and processes governing online speech' (2017) 131 *Harvard Law Review* 1598. Similarly, Emily Laidlaw has described the regulatory function of platforms as one of gatekeeping. Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press 2015).

7 Julie Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press 2019).

select and organise information as it is shown to users.[8] Recommender systems decide over visibility and virality in platform ecosystems, propelling some speakers to fame and relegating others to obscurity.[9] In a very literal sense, recommender systems define what it means to be relevant online.[10] As controversies proliferate over the role of social media in shaping our democracies—from political bias and racial discrimination to news diversity and quality—law and policy are now starting to hold platform recommender practices accountable to public values.[11]

In the push to regulate recommending, platforms have faced growing pressure to clarify how their recommender systems function. These systems are at present deeply opaque due to intentional corporate secrecy, as well as the sheer scale and complexity. Only platforms themselves hold all the data necessary to understand how they function. For users, governments and the public at large, this information asymmetry fuels anxiety and speculation. 'Dark ads' may be influencing elections without leaving a trace on the public record. 'Shadow bans' may be secretly censoring users without their knowledge. Some commentators have come to speak in nigh-mystical tones about 'The Algorithm' governing our online fates; powerful yet fundamentally unknowable.[12] But new legislation is attempting to open the black box, and regulate transparency in social media recommender systems.

Regulating transparency is by no means straightforward, especially not for complex algorithmic systems such as recommenders. Over the past decades, following many failed experiments, legal scholarship has come to acknowledge that transparency is no policy

8 Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' (2019) 10(3) *European Journal of Law and Technology* <https://ejlt.org/index.php/ejlt/article/view/686> accessed 15 September 2022.

9 The concept of media virality was coined by media theorist Douglas Rushkoff. More recent accounts have explored the role of platforms and their recommender systems in shaping virality. Douglas Rushkoff, *Media virus! hidden agendas in popular culture* (Random House 1996). Thomas Venturini, 'From fake to junk news: The data politics of online virality', in: Didier Bigo, Engin Isin and Evelyn Ruppert (eds.), *Data Politics: Worlds, subjects, rights* (Routledge 2019).

10 Tarleton Gillespie, 'The relevance of algorithms', in Tarleton Gillespie and others (eds), *Media Technologies: Essays on Communication, Materiality, and Society* (The MIT Press 2014).

11 José van Dijck, Thomas Poell and Martijn De Waal, *The Platform Society: Public Values in a Connective World* (Oxford University Press 2018.) Natali Helberger, Katarina Kleinen-Von Königslöw and Rob van der Noll, 'Regulating the new information intermediaries as gatekeepers of information diversity' (2015) 17 *info* 50.

12 Taina Bucher, 'The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms' (2017) 20 *Information, Communication & Society* 30. Simone Natale, 'Amazon can read your mind: A media archaeology of the algorithmic imaginary', in Simone Natale and Diana Pasulka (eds.), *Believing in Bits: Digital Media and the Supernatural* (Oxford University Press).

panacea, and more information does not guarantee accountability.[13] Designing effective disclosures requires a clear view of the topic at issue, the audience being addressed and the regulatory purpose served: transparency of what, transparency for whom, and to what end?[14] The transparency of machine-learning systems such as content recommenders is particularly challenging; their extreme complexity hinders attempts at straightforward explanation, and some have questioned whether the algorithmic transparency ideal of 'opening the black box' is at all feasible or meaningful.[15]

That is the challenge at issue in this dissertation. I have sought to describe how EU law regulates the transparency of recommender systems; to critique the regulatory functions that this transparency serves; and to explore how it can contribute to more democratic and inclusive social media governance. I have done so through five journal articles, compiled in this dissertation. What kinds of information is the law demanding about social media recommender systems? Who is included in these new models of accountability, and who is excluded? Focusing on social media, as opposed to other platforms, I am especially concerned with the ways that transparency can help to govern recommenders in light of public interest media principles. I aim to problematise the EU's technocratic reliance on regulators and trusted experts, and to open up social media transparency for more inclusive and overtly political aims.

This first chapter proceeds by introducing the main concepts under discussion: social media governance, recommender systems, and transparency and accountability. I then describe my research question, methodology and this dissertation's article-based format.

13 Mikkel Flyverbom, *The Digital Prism: Transparency and managed visibilities in a datafied world* (Cambridge University Press 2019). Mike Ananny and Kate Crawford, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2018) 20 *New Media & Society* 973.

14 Jakko Kemper and Daan Kolkman, 'Transparent to whom? No algorithmic accountability without a critical audience' (2019) 22 *Information, Communication & Society* 2081.

15 Bernhard Rieder and Jeanette Hofmann, 'Towards Platform Observability'(2020) 9(4) *Internet Policy Review* <https://doi.org/10.14763/2020.4.1535> accessed 19 September 2022. Lilian Edwards and Michael Veale, 'Slave to the algorithm? Why a "Right to an Explanation" is probably not the remedy you are looking for' (2017) 16 *Duke Law & Technology Review* 18.

2. Background: The problem of transparency in social media recommender systems

2.1 Governance *of* and *by* social media platforms

This dissertation starts from a concern with media governance and its platformisation by social media. Social media services are platform services, which is to say that they are internet-based digital services which host and organise user-generated content built on an infrastructure for processing data.[16] Platformisation, coined by Anne Helmond, refers to 'the rise of the platform as the dominant infrastructural and economic model of the social web and its consequences'.[17] In this emerging platform society, a growing number of societal domains have come to be (re)intermediated and (re)structured by platform services.[18] The economics of platformisation exhibit a strong tendency toward market concentration; a winners-takes-all dynamic which tends to result in only a handful of platforms becoming dominant within any given domain.[19] One of the earliest and most controversial domains of platformisation has been the media industry, which have come to depend to an ever greater extent on social media platforms.[20]

Social media platforms are the subset of platforms which revolve primarily around facilitating communications and social exchange, as distinct from other forms of interaction such as, for instance, retailing or transport.[21] Important examples include Facebook, YouTube, Instagram, and TikTok. Though there are substantial differences between each of these services, they have several features in common: social media platforms allow users to create profiles and publish content, such as text, image or video; to engage and interact with other profiles and content; and to navigate the available items through personalised recommender systems.[22] All major social media platforms are for-profit commercial corporations, and most of them are publicly listed. They earn revenue primarily through personalised advertising, which is targeted at individual users based on detailed personal data profiles assembled by tracking their online activity on and off the platform.[23]

16 Tarleton Gillespie, *Custodians of the internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018).

17 Anne Helmond, 'The platformization of the web: Making web data platform ready' (2015) 1(2) *Social Media + Society* <https://doi.org/10.1177/2056305115603080> accessed 19 September 2022.

18 Van Dijck, Poell, and De Waal (n 11), *The Platform Society*.

19 Martin Moore and Damian Tambini, *Digital dominance: The power of Google, Amazon, Facebook, and Apple* (Oxford University Press 2018).

20 Kleis Nielsen and Ganter, *The Power of Platforms* (n 3). Van Dijck, Poell and de Waal, *The Platform Society* (n 11).

21 Jean Burgess, Alice Marwick, A. and Thomas Poel (eds.), *The SAGE Handbook of Social Media* (Sage 2017).

22 Ibid.

23 Sophie Boerman, Sanne Kruikemeier and Frederik Zuiderveen Borgesius, 'Online Behavioral Advertising: A Literature Review and Research Agenda' (2017) 46 *Journal of Advertising* 363.

The largest social media platforms have become central fixtures of the contemporary media ecosystem. Many mass media organisations now maintain a presence on social media platforms; even though they compete with social media for advertising revenue, these media organisations also depend on social media as a means to reach online audiences, which represent a growing share of overall media consumption.[24] Van Dijck, Poell and de Waal summarise this transformation of media ecosystems in terms of their datafication (platform metrics like engagement come to dominate over, or even replace, conventional methods of professional editorial judgement), commodification (news is decontextualised and unbundled into standardised items of platform content) and selection (prominence is decided by the platform's complex and opaque algorithmic logics).[25] On social media these established media organisations are furthermore joined by new categories of online media actors; web-only outlets, professional content creators and 'influencers' as well as established public figures and politicians, all able to directly reach social media audiences without their past reliance on conventional media institutions.[26] In addition, as an interactive medium, social media platforms provide the occasion for ordinary users to interact directly with these outlets and with one another, integrating the private discourses of ordinary citizens with public discourses of the media in a novel semi-public digital environment.[27] In this new networked media environment, therefore, the interests and concerns of media governance are increasingly tied up with the fates of social media platforms and their governance.

How, then, are social media governed? Following Tarleton Gillespie, social media governance can be understood in terms of as a governance *of* and *by* platforms: platforms act as influential governors of social media ecosystems, and are in turn governed by law and other societal forces.[28]

The governance *by* platforms is described by Thomas Poell, David Nieborg and Erin Duffy as comprised of three main modalities: regulation, content curation, and

24 Kleis Nielsen and Ganter, *The Power of Platforms* (n 3).

25 Van Dijck, Poell and De Waal, *The Platform Society* (n 11).

26 Andrew Chadwick, *The Hybrid Media System: Politics and power* (Oxford University Press 2017).

27 e.g. Danah Boyd, 'Social network sites as networked publics: Affordances, dynamics, and implications', in: Zizi Papacharissi (ed.), *A Networked Self* (Routledge 2010). Thomas Poell, Sudha Rajagopalan and Anastasia Kavada, 'Publicness on platforms: Tracing the mutual articulation of platform architectures and user practices', in: Zizi Papacharissi (ed.), *A Networked Self and Platforms, Stories, Connections* (Routledge 2018).

28 Tarleton Gillespie, 'Governance of and by platforms', in: Jean Burgess, Alice Marwick and Thomas Poell (eds), *The SAGE handbook of social media* (Sage 2017).

content moderation.[29] Regulation describes the setting of standards and policies, both contractual and technical, which regulate how end-users and complementors (i.e. content providers, advertisers) can engage with the platform system.[30] Content curation describes the relative ordering or 'ranking' of content into navigable selections, carried out primarily through automated recommender systems.[31] Content moderation, finally, is the enforcement of rules applicable to user content and conduct.[32] Recommender systems, as I will explain further below, are primarily an instrument of content curation, but are also starting to be used as instruments of content moderation. Across all these modalities, it should be noted that the control which platforms exercise over their ecosystems is by no means absolute. Operating at massive scales and with limited knowledge of their users, platform strategies often struggle to predict or even measure the impact of their own designs. Efforts to classify users and content are inaccurate and frequently inflict collateral damage on legitimate users and behaviour.[33] Worse, these errors tend to disproportionally affect already-marginalised groups such as racial or linguistic minorities, since platforms rely on scalable machine-learning solutions which inherit real-world biases from their input data.[34] Conversely, platforms often fail to address many transgressions, and their enforcement strategies often encounter adaptation, resistance and circumvention from the users which they target.[35] In other words, users also exercise their own agency in platform governance; it emerges from the interactions between these user and the service.

The governance *of* social media platforms refers to the role of non-platform actors in

29 Thomas Poell, David Nieborg and Brooke Erin Duffy, *Platforms and cultural production* (John Wiley & Sons 2021).

30 Regulation by private entities, or governance, can arguably be difficult to distinguish from mere private action or coordination. I follow the work of Jeannette Hofmann, Christian Katzenbach and Kirsten Gollatz in conceiving of regulation as being defined by *reflexive* coordination, manifesting at critical junctures when established routines or practices generate conflict and lead to a (re)negotiation of the underlying norms, expectations and assumptions guiding conduct. See: Jeanette Hofmann, Christian Katzenbach and Kirsten Gollatz, 'Between coordination and regulation: Finding the governance in Internet governance' (2017) 19 *New Media & Society* 1406.

31 Kerstin Thorson and Chris Wells, 'Curated flows: A framework for mapping media exposure in the digital age' (2016) 26 *Communication Theory* 309.

32 Sarah Roberts, *Behind the Screen: Content moderation in the shadows of social media* (Yale University Press 2021).

33 Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) *Big Data & Society* 1 <https://doi.org/10.1177/2053951719897945> accessed 19 September 2022.

34 Ibid. Reuben Binns and others, 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation' (2017) *International Conference on social informatics* 405. Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press 2018).

35 Poell, Nieborg and Duffy, *Platforms and Cultural Production* (n 29).

shaping platform conduct. This governance includes legal but also non-legal factors. From a legal perspective, social media have historically been subjected to light-touch regulation known in the EU as the 'E-Commerce' framework.[36] This framework treats social media platforms as mere intermediaries, with only limited liability for the user-generated content which they host.[37] This relatively minimal programme is concerned mainly with combating unlawful publications whilst enabling cross-border service provision, and has little immediate regard for conventional principles of media policy such as pluralism, information quality, or access to news and educational material.[38] In recent years, however, the regulation of platforms has intensified as part of a broader societal 'techlash' starting approximately in 2016.[39] Many of these rules have continued to focus on the removal of unlawful content (such as terrorist propaganda and copyrighted works).[40] But, although media policy has remained primarily in the remit of the member states, still EU law and policy are starting to explore principles of media law and transpose these to platforms. An early steps in this direction is the revised Audiovisual Media Services Directive, which includes rules not only on illegal hate speech but also on child protection and advertising transparency.[41] The Code of Practice on Disinformation sets standards *inter alia* for social media fact-checking by platforms in cooperation with civil society partners.[42] Perhaps the most significant of these developments is the new Digital Services Act ('DSA'), which imposes duties on

36 Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (E-Commerce Directive).

37 E-Commerce Directive, Article 14.

38 See, generally: Olivier Castendyk, Egbert Dommering and Alexander Scheuer, *European Media Law* (Kluwer 2008). On the concept of 'public interest, see Philip Napoli, *Social media and the public interest: Media regulation in the disinformation age* (2019 Columbia University Press). Philip Napoli, *Foundations of communications policy: Principles and process in the regulation of electronic media* (Hampton Press 2001).

39 Ben Zimmer, 'Techlash': Whipping Up Criticism of the Top Tech Companies', *The Wall Street Journal* (10 January 2019) <https://www.wsj.com/articles/techlash-whipping-up-criticism-of-the-top-tech-companies-11547146279> accessed 24 September 2022.

40 Regulation 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (TERREG). Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Copyright Directive).

41 Directive 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services in view of changing market realities (AVMS Directive).

42 EU Code of Practice on Disinformation (European Commission 2018) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> accessed 27 September 2022. EU Strengthened Code of Practice on Disinformation (European Commission 2022) < https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation> accessed 27 September 2022.

large platforms to observe media governance principles such as media freedom, media pluralism, the protection of civic discourse and the protection of minors.[43] Through this novel domain of social media law, the governance *by* social media platforms is increasingly hemmed in by public rulemaking.

Still, platforms governance goes much further than what is required by law, and is shaped by non-legal political, economic and social forces. Such non-legal factors drive many platform policies, including commercial considerations such as brand management and user satisfaction as well as social and political considerations such as reputational and regulatory risk.[44] To take one important example, the vast majority of content moderation actions by platforms are voluntary: they are based not on legal categories such as copyright liabilities but rather on internal 'house rules' such as prohibitions on nudity, spam, or disinformation.[45] These voluntary policies can be understood as part of the platform's commercial service model, and in many cases also as a strategic response to political or reputational risks and pressures.[46] In this sense, as Natali Helberger argues, platforms must also be understood as political actors, whose policies exercise opinion power over online media ecosystems.[47] Platform governance requires a multistakeholder perspective, attentive not only to legal institutions but to the many other actors who together influence and regulate platform conduct, including users, political actors and civil society actors.[48] Such extra-legal relationships can also be formalised: partnerships with civil society organisations and other expert bodies are increasingly common for platforms, as is co-regulatory standard-setting with government agencies.[49] In practice, the line between legal and non-legal mechanisms can be blurry since these third parties also

43 Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services and amending Directive 2000/31/EC 2020 [COM/2020/825 final] (Digital Services Act).

44 Tarleton Gillespie, *Custodians of the Internet* (n 16).

45 Paddy Leerssen, 'Cut out by the middle man: the free speech implications of social media blocking and banning in the EU' (2015) 6 *JIPITEC* 99. Tarlach McGonagle, 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation', in: Giancarlo Frosio (ed.), *The Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

46 Gillespie, *Custodians of the Internet* (n 16).

47 Natali Helberger, 'The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power' (2022) 8 *Digital Journalism* 842.

48 Robert Gorwa, 'What is platform governance?' (2019) 22 *Information, Communication & Society* 6.

49 Robert Gorwa, 'The platform governance triangle: conceptualising the informal regulation of online content' (2019) 8(2) *Internet Policy Review* 2 <https://doi.org/10.14763/2019.2.1407> accessed 19 September 2022. Robyn Caplan, Networked Platform Governance: Reconciling Horizontals and Hierarchies in the Platform Era (Doctoral dissertation, Rutgers The State University of New Jersey, School of Graduate Studies 2021). Brenda Dvoskin, 'Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance' (2022) 67 *Villanova Law Review* 447.

interact with legal norms and institutions, for instance by litigating (or threatening to litigate) or by spurring regulatory reforms through acts of whistleblowing, evidence-gathering, awareness raising and so forth.[50] Platform policy thus navigates a complex constellation of legal, economic and political pressures, in which the role for public law is appreciable but by no means exhaustive.

In this dissertation, I approach social media governance from the multistakeholder perspective of 'cooperative responsibility', as outlined by Natali Helberger, Jo Pierson and Thomas Poell.[51] This model resists attempts to fix and isolate responsibility in any particular stakeholder—be it platforms, users, governments, or civil society—and instead views the attainment of public values in platform governance as a collective achievement requiring interaction and coordination between these groups, hence as a cooperative process. This emphasis on cooperation should not be misunderstood as ruling out binding government regulation; on the contrary, the authors expressly argue that binding government action is necessary in light of the failures of platform self-regulation. But what cooperative responsibility does stress, is that government's policy ought not merely to ordain public norms directly through law, but also to use law to create conditions for users and civil society actors to articulate their own preferences and thereby contribute to the collective realisation of public values.

Transparency is one important barrier to cooperative responsibility, since relevant societal actors often lack the information necessary for meaningful public deliberation about platform governance values.[52] This is especially true for platform recommender systems, which are central instruments of platform governance but remain shrouded in secrecy.

2.2 Recommender systems as points of control in social media governance

Until now, most legal scholarship on social media and user-generated content regulation has focused on platforms' hosting functions, and more specifically on whether and when platforms are required to remove unlawful content.[53] This

50 Margot Kaminski, 'Understanding Transparency in Algorithmic Accountability' in: Woodrow Barfield (ed.), *Cambridge Handbook of the Law of Algorithms* (Cambridge University Press 2020).

51 Natali Helberger, Jo Pierson and Thomas Poell, 'Governing online platforms: From contested to cooperative responsibility' (2018) 34 *The Information Society* 1.

52 Ibid.

53 e.g. Christina Angelopoulos, *European Intermediary Liability in Copyright: A Tort-Based Analysis: A Tort-Based Analysis* (Kluwer 2016). Martin Husovec, *Injunctions against intermediaries in the European Union: accountable but not liable?* (Cambridge University Press 2017). Jonathan Zittrain, 'A History of Online Gatekeeping' (2006) 19 *Harvard Journal of Law and Technology* 253. Michel Peguera, 'The DMCA Safe Harbors and Their European Counterparts: a comparative analysis of some common problems' (2009) 32 *Columbia Journal of Law & the Arts* 481.

dissertation instead addresses the comparatively novel regulation of recommender systems, which is rapidly gaining ground as a means to regulate lawful content without removing it completely.[54]

Recommender systems, per one technical definition, are 'software tools and techniques that provide suggestions for items that are most likely of interest to a particular user'.[55] In this dissertation, I use recommender systems in a broad sense which includes spontaneous recommenders, such as 'feeds' or 'trending' features, as well as search features which require users to enter a query before receiving suggestions. Important examples include Instagram's Feed, Facebook's Newsfeed, Youtube's Recommended Videos, and Tiktok's For You section. My analysis is limited to recommender systems operated by social media platforms themselves, and does not cover independent third party search engines such as those provided by Google Search or Google News, which conduct a comparable form of algorithmic ranking but have a different technical and legal relationships to the indexed content.[56] Although third party search engines exercise their own influence on media governance, this dissertation focuses on social media platforms for the sake of coherence and consistency.[57] For similar reasons, my analysis also excludes recommender systems operated by news organisations themselves, which they use optimise the offerings on their own content portals.[58] My discussion of social media recommender systems *does* include social media advertising functions, which are likewise supplied to users in an automated, personalised fashion and often within the same graphical interface as ordinary, so-called 'organic' content recommendations. These advertising functions can be understood as a form of paid priority within the recommender system.

54 Eric Goldman, 'Content Moderation Remedies' (2021) 28 *Michigan Technology Law Review* 1. Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' (2019) 10(3) *European Journal of Law and Technology* <https://ejlt.org/index.php/ejlt/article/view/686> accessed 15 September 2022.

55 Francesco Ricci, Lior Rokach and Bracha Shapira (eds.), *Recommender Systems Handbook* (Springer 2015).

56 Joris van Hoboken, *Search Engine Freedom: On the Implications of the Right to Freedom of Expression for the Legal Governance of Web Search Engines* (Kluwer International 2012). Wolfgang Schulz, Thorsten Held and Arne Laudien, 'Search Engines as Gatekeepers of Public Communication: Analysis of the German framework applicable to Internet search engines including media law and anti-trust law' (2005) 6 *German Law Journal* 1419.

57 Ibid.

58 Nick Diakopoulos, *Automating the News: How Algorithms are Rewriting the Media* (Harvard University Press 2019). Sarah Eskens, *The fundamental rights of news users: The legal groundwork for a personalised online news environment* (Doctoral Dissertation, University of Amsterdam, Faculty of Law 2021).

The purpose of recommender systems, as it is commonly described, is to surface relevant content for users. This concept of 'relevance' still leaves much room for interpretation.[59] Relevance, for recommender systems, is defined through an algorithmic filtering or ranking process, which can be relatively simple or deeply complex. For instance, a straightforward reverse chronology algorithm rank items according to their time of upload—in effect, equating relevance to timeliness. In practice, most social media platforms use machine-learning methods, which draw on large sets of users' personal data to make predictions about their content and behaviour.[60] The overarching commercial goal is typically to optimise for user attention and engagement (measured through proxies such as watch time, clicks, comments, or 'likes'), so as to maximise the possibility for ad placements. In turn, these ad placements are optimised based on targeting criteria supplied by the ad buyer and the platform's pricing system.[61] Subsequent chapters will describe the precise characteristics of these systems in greater detail. Here it suffices to note that recommender systems are not just instruments of commercial optimisation, but increasingly instruments of governance—which is to say that they increasingly take on reflexive considerations intending to steer the system toward specific types of conduct.[62] Taking a media governance perspective, I am especially interested in recommender systems' role in content regulation (as opposed to for instance, data protection's concern with the processing of personal data in recommender systems or competition law's concern with anti-competitive self-preferencing.)

As instruments of content regulation, recommender systems are implicated firstly in content curation and second in content moderation. First, as a matter of content curation, platforms constantly tweak and adapt how they measure and optimise for 'relevance', in ways which are designed not only to optimise for engagement but also, and increasingly, to accommodate political, legal and strategic considerations.[63] Second, as matter of content moderation, platforms increasingly intervene to sanction certain items via their recommender systems, as a means to enforce their house rules on content and conduct.[64] These 'visibility restrictions' include delisting, which removes the item from

59 Tarleton Gillespie, 'The relevance of algorithms', in Tarleton Gillespie and others (eds), *Media Technologies: Essays on Communication, Materiality, and Society* (The MIT Press 2014). Elizabeth van Couvering, 'Is relevance relevant? Market, science, and war: Discourses of search engine quality' (2007) 12 *Journal of Computer-Mediated Communication* 866.

60 Ricci, Rokach and Shapira, *Recommender Systems Handbook* (n 55).

61 Joseph Turow, *The Daily You: How the News Advertising Industry is Defining Your Identity and Your Worth* (Yale University Press 2011). Frederik Zuiderveen Borgesius, *Improving Privacy Protection in the Area of Behavioural Targeting* (Kluwer International 2015).

62 Several examples will be discussed in Chapters 2 and 5 below.

63 Napoli, *Social media and the public interest* (n 9).

64 Tarleton Gillespie, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 *Social Media+ Society* <https://doi.org/10.1177/205630512211175> accessed 19 September 2022.

a given recommender, and demotion (or downranking), which reduces its prominence relative to other items within the system.[65] Broadly speaking then, content curation is a positive selection for what *should* be visible or prominent on social media (content curation), whereas content moderation speaks to the negative selection of what *should not* be visible or prominent (content moderation).[66] In both cases, recommender governance must aim to strike a fair balance between competing media policy interests, both old and new, which I discuss below.

From a media policy perspective, content curation via recommender systems speaks first and foremost to questions of media diversity and quality. As platforms take up an increasingly important role in shaping exposure diversity, the composition of their offerings becomes a matter of increasing public interest. Recommender outcomes are therefore starting to be critiqued based on established media pluralism principles, such as the adequate furnishing of news content, educational material, representation of minorities and a diversity of political viewpoints.[67] In addition, the personalised nature of recommender outcomes raises new diversity concerns of their own, which are only starting to be articulated; recent debates around concepts such as 'filter bubbles', 'rabbit holes' and 'echo chambers' reflect a growing concern with the quality and diversity of individual media diets, and how these are shaped by recommender systems.[68] And yet, as Natali Helberger argues, the turn towards holding platforms responsible for content curation outcomes also creates tensions with the tradition of media concentration regulation, which reflects a concern with the dispersal of power over public discourse.[69] From this perspective, more proactive interference from platforms in content curation could actually be *harmful* to media pluralism, insofar as it normalises and reinforces their exercise of opinion power.[70] For Helberger, the challenge of contemporary media policy is therefore not merely to encourage platforms to adopt media diversity or other

65 Ibid.

66 Ibid. Per Gillespie, content curation selects *in*, and content moderation selects *out*.

67 Napoli, *Social media and the public interest* (n 9). Natali Helberger, Kari Karppinen and Lucia d'Acunto, 'Exposure diversity as a design principle for recommender systems' (2018) 21 *Information, Communication & Society* 191.

68 Technically, what distinguishes these concerns is as an attention for *vertical* diversity (i.e. within individual media diets) rather than horizontal diversity (i.e. across demographics). James Webster, 'Diversity of exposure', in Philip Napoli (ed), *Media Diversity and Localism* (Routledge 2007). Eli Pariser, *The Filter Bubble: What the internet is hiding from you* (Penguin 2011). Axel Bruns, *Are Filter Bubbles Real?* (Polity Press 2019). Frederik Zuiderveen Borgesius and others, 'Should we worry about filter bubbles?' (2016) 5(1) *Internet Policy Review* <https://doi.org/10.14763/2016.1.401> accessed 20 September 2020. Mark Ledwich and Anna Zaitsev, 'Algorithmic extremism: Examining YouTube's rabbit hole of radicalization' (2019) 25(3) *First Monday* <https://doi.org/10.5210/fm.v25i3.10419 > accessed 19 September 2022. O'Callaghan D and others, 'Down the (white) rabbit hole: The extreme right and online recommender systems' (2015) 33 *Social Science Computer Review* 459.

69 Helberger, 'The political power of platforms' (n 47).

70 Ibid.

public interest values, but also to devise new forms of counter-vailing power to act as a check on the systemic opinion power of platforms and their recommender systems, and to democratise control over them.[71] For these reasons, the extent of social media's public interest responsibilities remains subject to debate.

As regards their role in content moderation, recommender systems do not map clearly onto existing categories of media policy. Legal scholarship and policymaking on social media regulation until recently focused primarily on the removal of unlawful content, and, accordingly, the question of intermediary liability of platforms for hosting this unlawful content.[72] Visibility restrictions remained out of scope since they are directed at lawful content; they continue to host the item while reducing its prominence. For these reason, visibility restrictions are presented as a less restrictive alternative to removal, fulfilling an important role in managing content quality (e.g. by suppressing spam and clickbait) and helping to respond to new categories of 'lawful but awful' online harms such as disinformation and political extremism.[73] In this dissertation I will argue that this conventional account overlooks the deep opacity of visibility management, which arguably makes it *more* restrictive than conventional takedown methods.[74] In any case, as with conventional content moderation there are concerns about excess, error and bias. Legal scholarship has debated whether users might have fundamental rights safeguards against such excesses, but the positive law offers relatively little guidance to resolve such conflicts.[75] First, established doctrine is primarily concerned with state censorship rather than interferences by private platforms and the novel contexts and content categories which their moderators contend with.[76] Second, what few precedents we do have remain focused on the more severe cases of content removal and account suspension, rather than

71 Ibid.

72 e.g. Giancarlo Frosio (ed.), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020). Folkert Wilman, *The Responsibility of Online Intermediaries for Illegal User Content in the EU and the US* (Edward Elgar 2020). Martin Husovec, *Injunctions against Intermediaries in the European Union* (n 53). Christina Angelopoulos, *European Intermediary Liability in Copyright* (n 53).

73 Gillespie, 'Do Not Recommend' (n 64).

74 See Chapter 5 below.

75 Mattias Kettemann and Anna Sophia Tiedeke, 'Back up: Can users sue platforms to reinstate deleted content?' (2020) 9(2) *Internet Policy Review* <https://doi.org/10.14763/2020.2.1484> accessed 19 September 2022.

76 Evelyn Douek, 'The limits of international law in content moderation' (2021) 6 *UC Irvine Journal of International, Transantional and Comparative Law* 37. Leerssen, 'Cut out by the middle man' (n 45). Barrie Sander, 'Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation' (2019) 43 *Fordham International Law Journal* 939. Rachel Griffin, 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality', SSRN Draft Paper (2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064738> accessed 19 September 2022. Tarlach McGonagle, 'Free Expression and Internet Intermediaries' (n 45).

mere visibility reduction.[77] For these reasons, the legal status of visibility remedies is only now starting to register in legal debate around content moderation.

In sum, the governance of social media recommender systems demands a complex weighing of public interests, which has until recently remained largely unregulated by law. This is set to change with the new Digital Services Act (DSA), first proposed on 15 December 2020 and adopted on 4 October 2022.[78] Although this is by no means the first European legislation for social media, it is the first to address in any detail the working of their recommender systems. The DSA regulates their role in both moderation in curation. As regards moderation, it introduces a due process framework aimed at procedural fairness and consistency.[79] As regards curation, it introduces a systemic risk management framework which requires large platforms to mitigate risks posed by their recommender systems to fundamental rights and other public interests, including media policy principles such as media pluralism, the protection of minors and the protection of civic discourse.[80] In this way, recommender governance is gradually becoming a matter of legal concern, subject to public oversight by regulators and by courts. Still, the DSA is only a first step, introducing broad standards which remain to be specified in subsequent standard-setting.

In this dissertation, I approach the nascent domain of recommender governance from a primarily procedural perspective. I do not aim to develop a substantive vision for the appropriate design or operation of recommender systems, nor to prescribe how these should realise public interest media principles such as diversity or quality. Following agonistic accounts of (social) media governance from authors such as Kari Karppinen, Marijn Sax and Naomi Appelman, I view these as intrinsically political

77 Ketteman and Tiedeke, 'Back up' (n 75).

78 Regulation 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). At EU level, the Platform-To-Business (P2B) regulation preceded the DSA in regulating recommender systems. This instrument is discussed briefly in Chapters 2 and 5. Its provisions focus primarily on commercial fairness and transparency. It is not studied in detail since it does not address public interest principles related to media policy at issue in this dissertation, and because its relevant contents overlap substantially with those of the DSA whilst being narrower in scope. For instance, the P2B regulation contains several due process rights for business users with relevance for algorithmic ranking, laid down in Articles 3-5, but the DSA offers equivalent rights not only to business users but to all users. By focusing on the DSA, my discussion of due process rights in Chapter 5 focuses on the most germane and far-reaching of these two instruments. See: Regulation 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services ('P2B Regulation').

79 See Section 5.3 below.

80 See Chapter 6 below.

issues which are not solvable in any objective, universal or definitive manner.[81] Hence my goal in this project is not to resolve tensions and conflicts over the public interest in recommender governance, but rather to render them productive; to make them available to inclusive democratic contestation.[82] This approach also aligns with Helberger's emphases on opinion power and, with Pierson and Poell, cooperative responsibility, which invite us to ask not only how platforms ought to act but also who gets to decide on these norms, and how to institute counter-vailing powers which can curtail platforms' until now largely unfettered discretion.[83] For all these authors, and indeed for most authors in the platform governance literature, *transparency* therefore emerges as an important avenue for reform. As a precondition for practically all democratic reform projects, transparency seems to be a rare point of consensus in an otherwise contentious debate. And yet, transparency has a complex regulatory politics of its own.

2.3 Of what and for whom? Transparency in recommender systems

This dissertation is concerned with the principle of transparency in recommender governance. As a working definition, transparency can be defined for our purposes as 'the disclosure of certain information that may not previously have been visible or publicly available'.[84] Transparency is a core principle of modern governance theory, though its precise meanings and functions remain ambiguous and contested.[85] Broadly speaking, transparency is seen as instrumental to the achievement of accountability; or 'a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences.'[86] Transparency policies originate in 17th century Enlightenment principles of limited government, informed

81 Kari Karppinen, *Rethinking Media Pluralism* (Fordham University Press 2013). Marijn Sax, 'Algorithmic News Diversity and Democratic Theory: Adding Agonism to the Mix' (2022) *Digital Journalism* <https://doi.org/10.1080/21670811.2022.2114919> accessed 25 September 2022. Naomi Appelman, 'Algorithmic content moderation through an agonistic lens: contesting online exclusion', Paper presented at MANCEPT: Digital Democracy, Governance and Resistance in a Digital Era (7 September 2022.).

82 Ibid.

83 Helberger, Pierson and Poell, 'From contested to cooperative responsibility' (n 51). Helberger, 'The Political Power of Platforms' (n 47).

84 Oana Albu and Mikkel Flyverbom, 'Organizational transparency: Conceptualizations, conditions, and consequences' (2019) 58 *Business & Society* 268-297.

85 Christopher Hood and David Heald (eds.), *Transparency: The key to better governance?* (Oxford University Press for the British Academy 2006).

86 Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 *European Law Journal* 447.

citizenship and democratic self-rule.[87] Hence, transparency originally focused on public bodies, manifesting in policies such as open records laws, public proceedings, fiscal disclosures, and so forth.[88] In the 20th century, and especially after the neoliberal turn to global and corporate governance, powerful private entities have also faced increasing demands of transparency, leading to a proliferation of regulatory and self-regulatory public reporting standards, product labelling standards, and so forth.[89] In this way, transparency and accountability ideals have followed concentrations of power, first public and now private, reflecting the democratic principle that the exercise of power should be knowable to those it affects.[90]

Transparency may aim at accountability, but accountability comes in many forms. The influential work of Mark Bovens distinguishes at least four important categories of accountability mechanisms, or relationships: political (toward elected representatives, political parties, voters), administrative (towards auditors, inspectors, controllers), professional (towards professional peers), and social (toward interest groups and other stakeholders).[91] Like much of the literature on this topic, Bovens' taxonomy describes the accountability of *public* institutions. But transparency and accountability of private corporations such as platforms is somewhat different. First, corporations are subject to market discipline (to a greater or lesser extent), and a central goal of many corporate transparency reforms is to enable informed (consumer) economic choices and thus spur competition around the disclosed practices, as a form of commercial or market-based accountability.[92] Second, corporations are not directly subject to political accountability mechanisms such as elections or ministerial appointments. Corporations might nonetheless be sensitive to political pressures, to some greater or lesser extent, in the broader sense that they are (conditionally) responsive to shifts in reputation and public opinion. Still, under Bovens' taxonomy, this is best

87 Robert Gorwa and Timothy Garton Ash, 'Democratic Transparency in the Platform Society' in: Nate Persily and Joshua Tucker (eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge University Press 2020). Emmanuel Alloa, 'Transparency: A magic concept of modernity', in: Emmanuel Alloa and Dieter Thomä (eds.), *Transparency, society and subjectivity* (Palgrave Macmillan 2018).

88 Gorwa and Garton Ash, 'Democratic Transparency in the Platform Society' (n 87). Mark Warren, 'Accountability and democracy', in: Mark Bovens (ed), *The Oxford handbook of public accountability* (Oxford University Press 2014).

89 Radu Mares, 'Corporate transparency laws: A hollow victory?' (2018) 36 *Netherlands Quarterly of Human Rights* 189. David Pozen observes that corporate transparency regulation is at least as old as the US progressive era. David Pozen, 'Transparency's Ideological Drift' (2018) 126 *Yale Law Journal* 100.

90 Warren, 'Accountability and democracy' (n 88).

91 Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 *European Law Journal* 447.

92 Archon Fung, David Weil and Mary Graham, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge University Press 2007).

understood as 'social accountability', as it plays out through civil society and public debate rather than through formalised procedures or forums. Given this diversity of different possible mechanisms, accountability is not so much a singular substantive policy concern as it is a language for describing power relationships and disciplinary mechanisms.[93] Accountability, for Jerry Mashaw, is a 'protean' concern seeking to constrain the conduct of powerful organisations but then begging the question: accountability of what conduct? And accountability to whom?[94] These underlying questions also inform transparency policy.

It is often observed that transparency is not a guarantee of accountability, though it may be a precondition. Other conditions must also be met. Even when transparency brings to light wrongdoing, the relevant stakeholders must still be willing and able to make use of this information and sanction transgressions. In other words, it requires power. For law and policy, an important question is therefore how transparency relates to behavioural regulation; should transparency serve legal or non-legal forms of accountability? In other words, should transparency be accompanied by behavioural regulation, or act as an alternative to it?[95] Over the past decades, also in digital policy, transparency has typically been conceived of as a market-based intervention, oriented around individual consumer choice rather than public rulemaking and therefore associated with a neoliberal politics of deregulation.[96] This model has come under extensive criticism, with the growing recognition that individual consumers often lack the wherewithal to process transparency disclosures and/or the power to act on these disclosures.[97] More recent commentary on transparency therefore emphasises the importance of coupling it to a substantive programme of behavioural regulation and strengthening the regulatory institutions entrusted with this task.[98] How, precisely, transparency can contribute to regulation has not received as much detailed attention, and one of the goals of this dissertation will to be to unpack this complex relationships in the specific context of social media recommender systems.

93 Warren, 'Accountability and democracy' (n 88).

94 Jerry Mashaw, 'Accountability and Institutional Design: Some Thoughts on the Grammar of Governance', (2006) Yale Law School Public Law Working Paper No. 116. See also: Ananny and Crawford, 'Seeing without knowing' (n 13).

95 Monika Zalnierute, '"Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism' (2021) 8 *Critical Analysis of Law* 39.

96 Ananny and Crawford, Seeing without knowing (n 13).

97 Frederik Zuiderveen Borgesius, 'Behavioural sciences and the regulation of privacy on the internet', in: Anne-Lise Sibony and Alberto Alemanno (eds.), *Nudging and the law: what can EU law learn from behavioural sciences?* (Hart Publishing 2015).

98 Pozen, 'Transparency's Ideological Drift' (n 89).

For all its ambiguities, the principle of transparency is distinctly relevant to the case of platforms, and in particular to their recommender systems. Platform governance is characterised by deep information asymmetries, with the platform itself possessing vast troves of granular data about its service, which users and third parties can access only partially and selectively.[99] Indeed, this information asymmetry is part and parcel of the platform's informational capitalist business model, which aims to collect and monetise user data as an exclusive asset.[100] Besides economic self-interest, platforms also appeal to user privacy and service security as reasons to maintain confidentiality. Recommender systems in particular are criticised for their opacity. Even if platforms were to disclose their underlying algorithms, the machine-learning techniques used to rank content are so complex and so unlike human cognition that they resist explanation in ways that are understandable to non-experts.[101] For all these reasons, meaningful transparency for recommender systems is by no means straightforward and may require very different designs tailored to the specific needs and capacities of different stakeholder groups.

This dissertation examines the role of EU law in bringing about transparency of social media recommender governance. As mentioned, this comprises a dual question: transparency of what, and transparency for whom. What aspects of recommender systems is the law bringing into view, and what remains occluded? What types of accountability relationships do these transparency reforms foresee, and under what conditions might they take hold? This dissertation develops two strands of critique as to recent policymaking on this issue, corresponding to these two components of transparency. As to the substance of recommender transparency (transparency of what?), I will argue for a sociotechnical perspective which moves beyond algorithmic explanations as the sole or primary topic of interest. As to the addressees of recommender transparency (transparency for whom?), I will argue for an inclusive, scalable approach emphasising public resources and civil society in the broadest sense.

As to substance, a sociotechnical perspective on recommender systems decenters the 'algorithm' as sole object of scrutiny. Until now, most scholarly work on transparency in automated decision-making has focused on the challenges of algorithmic transparency, and in particular devising new techniques for 'opening the black box' by producing understandable and salient explanations for their machine-learning

99 e.g. Pozen, 'Transparency's Ideological Drift' (n 89). Zalnierute, '"Transparency Washing" in the Digital Age' (n 95).

100 Cohen, *Between truth and power* (n 7).

101 Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2017) 3(1) *Big Data & Society* <https://doi.org/10.1177/2053951715622512> accessed 15 September 2022.

decisions.[102] However, I aim to demonstrate that the transparency problem for recommender systems goes further than algorithms alone; recommender systems are also opaque in other important ways, pertaining not just to their algorithmic logics but to their inputs, their outputs, and their organisational embedding; in other words, the ways they find application in social practice.[103] To this end I draw on work in critical algorithm studies, which approaches recommender systems not only as algorithmic artefacts but as sociotechnical systems, defined not only by their technical parameters but how these find use in social practice.[104] This sociotechnical perspective leads me to analyse recommender systems in terms of the distinct governance functions they fulfil in practice: content curation and moderation. I aim to show how this opens up new opportunities for transparency beyond the algorithm.

As to the addressees, I am interested in exploring how transparency measures include certain stakeholder groups and exclude others—how they prefigure and reinforce accountability relationships and regulatory structures. If it is agreed that transparency ought to support 'regulation', then the complex multistakeholder arrangements of platform governance still beg the question as to who is involved in such regulation—and, hence, who should be entitled to the benefits of transparency. Based on the principles of cooperative responsibility, my goal is to explore how transparency can assist not only in top-down government standard-setting but also in a more broad-

102 e.g. Karen Yeung, 'Algorithmic regulation: A critical interrogation' (2018) 12 *Regulation & Governance* 505. Frank Pasquale, *The Black Box Society: The secret algorithms that control money and information* (Harvard University Press 2015). Burrell, 'How the machine thinks' (n 101). Bryce Goodman and Seth Flaxman, 'European Union regulations on algorithmic decision-making and a "right to explanation"' (2017) 38 *AI Magazine* 50. Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31 *Harvard Journal of Law and Technology* 841. Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 2. But c.f. Edwards and Veale, 'Slave to the algorithm?' (n 15)(arguing that' a right to an explanation is probably not the remedy you are looking for'); and Margot Kaminski, 'The Right to Explanation, Explained' (2019) 34 *Berkeley Technology Law Journal* 189 (arguing that the 'the recent debate over the "right to explanation" […] obscured the significant algorithmic accountability regime established by the GDPR').

103 This sociotechnical perspective is a foundational theme in Science and Technology Studies (STS). See e.g. Eric Trist and Ken Bamforth, 'Some social and psychological consequences of the long-wall method of coal-getting', (1951) 4 *Human Relations* 3. Bruno Latour, *Science in Action* (Harvard University Press 1987). Wiebe Bijker, *Of Bicycles, Bakelites, and Bulbs* (MIT Press 1995). Langdon Winner, 'Do Artifacts Have Politics?', (1980) 109 *Daedalus* 177.

104 Ananny and Crawford, 'Seeing without knowing'. Nick Seaver, 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems' (2017) 4(2) *Big Data & Society*. <https://doi.org/10.1177/2053951717738104> accessed 19 September 2022. Mike Ananny, 'Toward an ethics of algorithms: Convening, observation, probability, and timeliness' (2017) 41 *Science, Technology, & Human Values* 93. Rieder and Hofmann, 'Towards Platform Observability' (n 15).

based and civil-society driven societal conversation around public values in platform governance.[105] The case of social media, as distinct from other platforms, is particularly salient due to their unique role in democracy as a site of political power.[106] Media policy is typified by a distinct suspicion of direct government ordering, and places special emphasis on the importance of inclusive and egalitarian governance.[107] And yet, due to sensitivity of platform data, this dissertation will show, transparency policy is inclined toward confidential and technocratic modes of governance, with data being made available only to a trusted few at government agencies and research institutions. As a counterweight to these tendencies, this dissertation argues for a more nuanced appreciation of the importance of public and broadly inclusive forms of data access, as a precondition for democratically accountable recommender governance.

3. Research question and outline

This dissertation aims to answer the following research question:

> How can EU law regulate the transparency of recommender systems in order to hold online platforms accountable for their role in social media governance?

I address this question by way of the following sub-questions:

1. What different models of accountability are reflected in the EU's regulation of recommender system transparency for social media?
2. How can transparency regulation contribute to the accountability of content curation through recommender systems?
3. How can transparency regulation contribute to the accountability of content moderation through recommender systems?
4. What is the relationship between transparency regulation and behavioural regulation in social media recommender governance?

Chapter 2 addresses the first sub-question. After introducing the basic technical and political-economic characteristics of social media recommender systems, this paper reviews various models for transparency and accountability under development in EU policymaking. It identifies three types of disclosure rules: individual disclaimers, regulatory audits and researcher access. On this basis it articulates an initial statement

105 Helberger, Pierson, and Poell, 'From contested to cooperative responsibility' (n 51).
106 Eric Barendt, *Freedom of speech* (Oxford University Press 2005).
107 Edwin C Baker, *Media concentration and democracy: Why ownership matters* (Cambridge University Press 2006).

of this dissertation's main normative positions: First, transparency should not focuses solely on recommender algorithms, but take a broader perspective of recommenders as sociotechnical systems. Second, mechanisms for social accountability should, inasmuch as possible, aim to realise inclusive public resources (as opposed to exclusive, confidential resources).

Chapters 3 and 4 both address the second sub-question. They do so by way of an in-depth case study of a novel transparency technique in platform self-regulation: platform ad archives. These tools offer public, machine-readable overviews of advertising distribution via major platform services, along with metadata on their origin and distribution. Ad archives make for a relevant case study since they reflect a sociotechnical and inclusive approach to transparency, documenting systemic outputs rather than algorithms and being made accessible to all.

This case study unfolds over two chapters. Chapter 3 introduces the phenomenon of ad archives from a governance perspective, describing their legal backgrounds and possible accountability functions, as well as the shortcomings in their current self-regulatory implementations. As public tools, ad archives have the potential to contribute to accountability in several ways. Their role in governance, I argue, includes potential legal effects, through regulatory monitoring and enforcement, but also social and discursive effects based on the capacity for journalists, academics and other civil society actors to more effectively respond to personalised campaigns. On this basis I provide several proposals as to how public regulation might aim to improve ad archives. Chapter 4 complements this theoretical account with an original empirical investigation into the usage of ad archives by journalists, including content analysis of journalistic outputs and in-depth interviews with relevant journalists. This research confirms that journalists have made repeated use of the ad archive for reporting purposes, providing evidence of both social and legal accountability. This further strengthens the case for regulation, confirming civil society's demand for data about content curation as well as their present reliance on platforms to acquire it.

Chapter 5 addresses the third sub-question, turning to the role of recommender systems in content moderation. Content moderation sanctions imposed via recommender systems, I argue, are less transparent than conventional methods such as content takedown or account suspension. To this end I draw on social science research into user experiences of 'shadow banning', which refers to moderation sanctions imposed in secret. Shadow banning, I argue, is made possible by the volatile and personalised dynamics of platform recommender systems, which serve to obscure visibility sanctions and insulate them from legal and social accountability. This

phenomenon speaks to the importance of transparent *outputs* in content moderation, as a complement to and precondition for algorithmic explanations. I then analyse the content moderation transparency rights laid down in the DSA, which I interpret as prohibiting shadow bans as part of its individual due process framework. In implementing these notice rights, I argue, an important challenge will be to define visibility restrictions as a category of moderation sanctions distinct from the routine operations of recommender curation.

Chapter 6 address the fifth and final subquestion, reflecting on the regulatory politics of transparency in social media recommender systems. To this end, it draws on the concept of 'observability', recently proposed by Bernhard Rieder and Jeanette Hofmann as a pragmatic, sociotechnically informed alternative to transparency which expressly aims to serve as a 'companion to regulation'.[108] Reviewing the DSA's rules on observability, I show how these might contribute to regulation of social media recommenders. In doing so I argue that the law should work toward a more nuanced appraisal of observability's regulatory functions, not only as an instrument of compliance monitoring and enforcement but also as a resource for knowledge production and public discourse.

4. Methods and Format

4.1 Methods

This dissertation is first and foremost a work of normative legal scholarship, but it also draws extensively on non-legal social sciences and humanities in the multidisciplinary field of social media studies. Below I describe my methods. I do so in some detail in order to make it comprehensible to both legal and non-legal audiences.

The legal research method is an interpretive, hermeneutic method which draws on authoritative legal sources such as legal statutes and court decisions to describe, systematise and critique the positive law—i.e. the law as it is currently prescribed by authoritative sources.[109] In order to construct legal doctrine, the legal method interprets legal materials through the same 'internal' perspective of legal practitioners such as judges.[110] In terms of its jurisdiction, my legal analysis is focused on the law

108 Rieder and Hofmann, 'Towards Platform Observability' (n 15).

109 Sanne Taekema, 'Theoretical and normative frameworks for legal research: Putting theory into practice' (2018) *Law and Method* <https://doi.org10.5553/REM/.000031> accessed 19 September 2022.

110 Leslie Green and Thomas Adams, 'Legal Positivism', *The Stanford Encyclopedia of Philosophy* (2019) <https://plato.stanford.edu/archives/win2019/entries/legal-positivism/> accessed 25 September 2022.

of the European Union. In terms of substance, it is focused on the emerging field of social media law, and in particular the DSA. One challenge for this project has been that the DSA was not proposed until the penultimate year of research, and finalised only months before its conclusion. The earlier sections therefore focus to a larger extent on preliminary policy reports and proposals at the EU and national level, in order to discern and critique the direction of lawmaking, now culminating in the DSA.

Since my project is concerned with platform governance, I also take into account rulemaking carried out *by* platforms through private ordering, self-regulation, and co-regulation.[111] For instance, platform Terms of Service contracts are an important component of applicable law in this space. By combining public law and platform practice, I am able to trace the interplay between these spheres of action, and the 'hybrid power' which emerges from the interaction of platform power and state power.[112] For instance, the practice of ad archiving which is central to Chapters 3 and 4 was technically voluntary during the time of writing, and yet I show that platforms instituted the practice only after governments first *proposed* legislation to similar effect. Such public-private interactions 'in the shadow of the state' have also resulted in the adoption of law-like co-regulatory policy instruments, such as the EU Code of Practice on Disinformation.[113] Policies such as these do not have the status of law from an internal legal perspective, and yet they exercise law-like regulatory functions which are crucial for our understanding of social media rulemaking from a governance perspective.

For both its factual and normative orientation this dissertation draws heavily on communications and media studies, and in particular the multidisciplinary fields of social media studies and algorithm studies. Factually, I rely on these fields to

111 On this distinction, see: Chris Marsden, *Internet co-regulation: European law, regulatory governance and legitimacy in cyberspace* (Cambridge University Press 2011).

112 Martin Fertmann and others, 'Hybrid institutions for disinformation governance: Between imaginative and imaginary', *Internet Policy Review* (16 May 2022) <https://policyreview.info/articles/news/hybrid-institutions-disinformation-governance-between-imaginative-and-imaginary/1669> accessed 25 September 2022. David Levi-Faur, 'Regulation and Regulatory Governance', in: David Levi-Faur (ed.), *Handbook on the Politics of Regulation* (Edward Elgar 2011). Michael Birnhack and Niva Elkin-Koren, 'The Invisible Handshake: The Reemergence of the State in the Digital Environment' (2003) 8(6) *Virginia Journal of Law & Technology* < https://law.bepress.com/taulwps/art54/> accessed 25 September 2022.

113 Hannah Bloch-Wehba, 'Global platform governance: private power in the shadow of the state' (2019) 72 *SMU Law Review* 27. The original Code of Practice was presented as self-regulatory, but the close involvement of the European Commission in enacting and implementing this instrument leads Aleksandra Kuczerawy to conclude that it is better characterised as co-regulatory. A subsequent 2022 Strengthened Code is more expressly co-regulatory in its design. Aleksandra Kuczerawy, 'Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?', in: Elzbieta Kużelewska and others (eds.), *Disinformation and Digital Media as a Challenge for Democracy* (Intersentia 2021).

characterise the technical and political-economic qualities of social media services and their recommender systems. As mentioned, my analysis is informed in particular by STS-based accounts which analyse platform recommenders from a sociotechnical perspective. Social media studies scholarship is also relevant to my work in a second capacity: as a *stakeholder* in platform transparency. As mentioned, researchers' knowledge production feeds into legal and social accountability. Accordingly, methodological considerations from social media studies offers me a window onto the technical, legal and ethical obstacles which civil society actors encounter in studying platforms and their recommender systems.[114] For these reasons, I am interested in the regulatory function exercised by social media studies research. Furthermore, in Chapter 4 I conduct my own empirical research: semi-structured interviews and content analysis into journalistic usage of Facebook's ad archive. In this chapter, the dissertation moves from cross-disciplinary borrowing of insights to a more integrated interdisciplinary sociolegal approach, deploying social science methods to answer empirical questions with relevance to both disciplines.

Going beyond mere description, normative legal research also aims to evaluate the law and propose improvements to it. To this end, normative scholarship requires a normative framework, which explicates the values or norms on which recommendations are based, so as to make these available for reasoned argumentation and disagreement.[115] This dissertation's normative framework is grounded in the basic principles of EU law, and informed by media theory scholarship.[116] In particular, as outlined in Section 2.2 above, my research starts from a concern with the realisation of public interest media principles in social media recommender governance, which are expressed in the DSA and can be traced back to constitutional

114 e.g. Axel Bruns, 'After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research' (2019) 22 *Information, Communication & Society* 1544. Richard Rogers, 'Social media research after the fake news debacle' (2018) 11 *Partecipazione e conflitto* 557.

115 'Sanne Taekema, 'Theoretical and normative frameworks for legal research: Putting theory into practice' (2018) Law and Method <https://doi.org10.5553/REM/.000031> accessed 19 September 2022.

116 Following Taekema, my normative stance does not distinguish strictly between an internal or external perspective, but instead aims at a pragmatic assessment of the law's capacity, in views of its social context, at realizing its stated goals. This pragmatic assessment is 'not limited by the posited values and principles in basic legal documents but can criticize beyond that positive content, using aspects of the normative theory that are tied to the values as formulated within positive law but which reach beyond that.' This approach does not attempt the possible influence of personal convictions in the construction and critique of legal doctrine, but instead aims to explicate those convictions inasmuch as possible.

principles of information freedom and media pluralism.[117] Still, the positive law on these fundamental rights offers policymakers comparatively little guidance as an evaluative criterion for recommender governance.[118] I have found no case law on media pluralism policy as regards recommender governance (even in the mass media context the jurisprudence on these positive dimensions of information freedom is relatively sparse), and the DSA's provisions, as we will see, remain unspecific as well.[119] This is not fatal to my project due to its focus on transparency, as a procedural condition for legitimate governance of these underlying substantive issues, through cooperative responsibility.[120] What I take from media policy, then, is a general concern for the egalitarian distribution of media power, and a general distrust of both solely commercial and solely governmental ordering of the public sphere.[121] The normative goal of this project, therefore, is to study how transparency can act as a check not only on the power of platforms vis-à-vis governments, but in a more general sense to act as a check on the distribution of power in this governance system as a whole, and on the hybrid power that emerges from the interactions between platform and state.

4.2 Format

A brief comment on the dissertation's article-based format is in order, since legal dissertations have traditionally taken the form of a single, book-length treatise. Article-based formats are still relatively new and atypical in this discipline. This dissertation is comprised of five main sections, which correspond to five articles of my writing submitted to peer-reviewed academic journals. At the time of writing this introduction, three have been published, and two are under consideration. The introduction and conclusion accompany these works as part of the dissertation, and are not intended to be published independently. For Chapters 3 and 4 I share co-authorship with colleagues from my research team, in keeping with the conventions of social science. Still, in all cases the text in this dissertation is solely written by my hand and my co-authors contributed in the conceptualisation, research and reviewing stages.[122]

The article-based approach I consider appropriate for the fast-changing topic at hand. It has allowed me to engage with on-going and time-sensitive debates in EU

117 DSA, Articles 14(4) and 34(1)(b). Helberger, Kleinen-Von Königslöw and Van der Noll, 'Regulating the new information intermediaries as gatekeepers of information diversity'. *Informationsverein Lentia and others v Austria* [1993] ECtHR 13914/88. *Jersild v Denmark* [1994] ECtHR 15890/89. *Verein Gegen Tierfabriken v Switzerland* [2001] ECtHR 24699/94. *Appleby and Others v United Kingdom* [2003] ECtHR 44306/98. but c.f. *Delfi v Estonia* [2015] ECtHR 64569/09.

118 Ibid.

119 Ibid.

120 Helberger, Pierson and Poell, 'Cooperative Responsibility' (n 51).

121 Edwin C Baker, *Media concentration and democracy* (n 107).

122 See also the Author Contributions statement appended to this manuscript.

digital policy, and to critique lawmaking during and not only after the fact. I will note that Chapter 2 of this dissertation is cited in the European Commission's Impact Assessment accompanying the DSA—not to self-aggrandise, but merely to drive home the importance of timely publication on this fast-moving issue.[123] I wish to underscore these considerations because the article-based approach also has its own drawbacks, for which I ask the reader's understanding. First, each article must be made to stand on its own as a separate publication, and this results in a degree of repetition and overlap between them. Second, published articles cannot be adjusted or updated over time to reflect, as a manuscript in progress might be, developments in law and scholarship, and the project's own thought and terminology. In accordance with faculty regulations the articles are reproduced here in their original form, as published, except for standardised orthography and referencing. Beyond this, only a handful of minor adjustments have been made to improve terminological clarity across articles, and each is specified in an accompanying footnote.

123 European Commission, Impact Assessment of the Digital Services Act (15 December 2020) <https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act> accessed 27 September 2022. Part I, Fn 54.

CHAPTER 2

The soap box as a black box: Regulating transparency in social media recommender systems[1]

1 Originally published as: Paddy Leerssen, 'The soap box as a black box: regulating transparency in social media recommender systems' 11(2) (2020) *European Journal of Law and Technology* <https://ejlt.org/index.php/ejlt/article/view/786>.

Abstract: *Social media recommender systems play a central role in determining what content is seen online, and what remains hidden. As a point of control for media governance, they are subject to intense controversy and, increasingly, regulation by European policymakers. A recurring theme in such efforts is transparency, but this is an ambiguous concept that can be implemented in various ways depending on the types of accountability one envisages. This paper maps and critiques the various efforts at regulating social media recommendation transparency in Europe, and the types of accountability they pursue.*

This paper identifies three different categories of disclosure rules in recent policymaking: (1) user-facing disclaimers, (2) government auditing, and (3) data-sharing partnerships with academia and civil society. Despite their limitations and pitfalls, it is argued, each of these approaches has a potential added value for media governance as part of a tiered, varied landscape of transparency rules. However, an important element is missing: public data access. Current trends emphasise exclusive data access regimes directed at particular, trusted regulators or researchers, but this approach has important limitations in terms of scalability, inclusiveness, and independence. This paper articulates the distinct benefits of public data access as a supplement to existing transparency measures, and suggests starting points for its design and regulation.

1. Introduction

Social media platforms have become central actors in media governance. One of their most powerful means of influence is their content recommender systems, which determine the ranking of content as it is presented to users. Their design can therefore have significant effects on what is seen online, and what remains hidden. Accordingly, content recommender systems have a gatekeeping function, implicating urgent public interests and swiftly becoming a key point of control and contention in ongoing debates about online content regulation.[2]

In this otherwise contentious debate, a rare point of consensus for both scholars and policymakers appears to be the need for greater transparency. At present, social media recommendation systems operate largely as 'black boxes', guided by complex, confidential machine-learning algorithms whose operations are inscrutable to outside observers.[3] 'A system must be understood to be governed', as Mike Ananny and Kate Crawford observe, and there is broad agreement amongst scholars and policymakers that recommender systems must be more transparent—if not as a sufficient condition for holding them accountable then at least as a first step.[4]

This paper analyses recent policymaking in Europe that attempts to regulate transparency in social media content recommendations. Not yet a cohesive framework, we see various overlapping standards at the national, EU and Council of Europe level, each furthering particular visions of 'transparency' and the types of accountability it should serve. This paper critiques these various efforts, drawing on critical literature on transparency regulation and platform governance, and questions how, and under which conditions, they can contribute to holding social media recommender systems accountable for their impact on online information flows.

The paper proceeds as follows: Section 2 offers an overview of governance debates about gatekeeping through social media recommenders, and how these have given rise to calls for greater transparency. Building on recent literature from communications and media studies, it is argued that transparency in social media recommendations is a multifaceted issue which relates not only to the algorithm involved but more

2 Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' (2019) 10(3) *European Journal of Law and Technology* https://ejlt.org/index.php/ejlt/article/view/686 accessed 15 September 2022.

3 Frank Pasquale, *The Black Box Society: The secret algorithms that control money and information* (Harvard University Press 2015).

4 Mike Ananny and Kate Crawford, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2018) 20 *New Media & Society* 973.

broadly to the design and operation of content recommenders as sociotechnical systems. Section 3 describes recent European policymaking around recommendation transparency, identifying three general categories of disclosure rules: user-facing disclaimers, government oversight and civil society partnerships. Each of these methods can have an added value for media governance, despite their respective limitations and pitfalls, as part of a tiered, variegated approach to transparency. Yet, an important element is missing: public disclosures. Section 4 articulates the distinct advantages of public disclosures as a supplement to existing transparency measures, and suggests starting points for their design and regulation.

2. Social media recommenders systems as opaque gatekeepers of online content

2.1 What are social media recommender systems?

Platforms use recommender systems to determine the manner in which content is presented to their users. Their recommendations typically take the form of pages or lists, often referred to as 'feeds', in which the order of content is determined by ranking algorithms. These ranking algorithms can take any number of forms, from simple reverse chronology to complex machine-learning solutions. Recommender systems can also include user customisation options, such as the ability to 'like' or 'follow' specific content sources, to block or filter certain content sources, or to switch between entirely different ranking logics. Recommender systems are commonly understood as optimising for user attention, or 'relevance', but in practice, as will be unpacked further below, recommender design is also shaped by other economic and political imperatives.

Recommendations are not the only way to access social media content. Users can typically also reach content through search functions, user profiles, hotlinking and embedding. Nonetheless, recommender systems can be highly influential, since they commonly take up a central position in platform interfaces: Facebook's Newsfeed and YouTube's Autoplay and Recommended Videos, for instance, are some of the key content discovery features on their respective platforms. YouTube recently stated that 70% of user viewing is accessed through recommendations.[5] Facebook's Newsfeed being even more central to the platform's interface, the percentage here could plausibly be even higher.

5 Karen Hao, 'YouTube is experimenting with ways to make its algorithm even more addictive', MIT Technology Review (27 September 2019) <https://www.technologyreview.com/s/614432/youtube-algorithm-gets-more-addictive/> accessed 17 September 2022.

It is important to note that content recommendations are not fully controlled by their operators, but are co-determined by platform users, who influence outcomes in several ways. Firstly, users are responsible for uploading content from which recommendations are generated. Secondly, users' behaviour provides feedback signals, including explicit feedback such as rating, following or subscribing, as well as implicit feedback such as scrolling and clicking.[6] Since recommender systems commonly rely on machine-learning processes to optimise the algorithm, these user signals can also serve to shape the weighting of the algorithm over time. Conversely, the recommender system can also shape users' behaviour over time, in terms of their preferences, habits and expectations they form in relation to the service. These complex interactions between the recommendation algorithm and its users make for a recursive and unpredictable system, with the potential for unexpected feedback loops and path dependencies. A notable example is Rebecca Lewis' study of far-right content on YouTube which emphasises the role of well-organised 'influence networks' of content creators and audiences, who used guest appearances and other forms of referral and collaboration to create a pipeline or 'rabbit hole' of gradually escalating extremism.[7]

Due to the central role of user behaviour in steering recommendation outcomes, platform recommendations are not fully pre-determined or controlled by their operators. For this reason, communications research into recommender systems has emphasised the importance of looking past algorithms as such towards understanding the complex interactions between technology and its users. Kevin Munger and Joseph Philips warn that decontextualised or monocausal understandings of 'the algorithm' shaping online media consumption overestimates the role of their designers and undervalues the relative influence of user communities which shape content supply and demand.[8] Philip Napoli instead characterises gatekeeping on social media as a process of 'individual media users working in conjunction with content recommendation algorithms'.[9] Building on such insights, Bernhard Rieder, Ariadna Matamoros-Fernandez and Oscar Coromina argue for a shift from studying ranking

6 Charu Aggarwal, *Recommender Systems: The textbook* (Springer 2016). Eytan Bakshy, Solomon Messing and Lada Adamic, 'Exposure to ideologically diverse news and opinion on Facebook' (2015) 348 *Science* 6239. David Lumb, 'Why scientists are upset about the Facebook Filter Bubble story', *Fast Company* (5 August 2015) <http://www.fastcompany.com/3046111/fast-feed/why-scientists-are-upset-over-the-facebook-filter-bubble-study> accessed 19 September 2022.

7 Rebecca Lewis, 'Alternative Influence: Broadcasting the Reactionary Right on YouTube' (Data & Society Research Report 2018) <https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf > accessed 19 September 2022.

8 Kevin Munger and Joshua Philips, 'Right-Wing YouTube: A Supply and Demand Perspective' (2020) 27 *The International Journal of Press/Politics* 186.

9 Philip Napoli, *Social media and the public interest: Media regulation in the disinformation age* (2019 Columbia University Press).

algorithms to 'ranking cultures', acknowledging 'the realities of an intricate mesh of mutually constitutive agencies'.[10] A more complete understanding of social media recommendations, then, cannot focus on recommendation algorithms alone but must seek to understand the sociotechnical system through which they are produced.[11] Rather, as Natali Helberger, Katharina Kleinen-Vön Königslöw and Rob von der Noll observe, governance of these systems therefore requires close attention to 'the complex dynamics between the gatekeepers and the gated'.[12] What this sociotechnical perspective demands in terms of transparency will be explored in section 2.3 below, but not before discussing the political economy of recommender system governance that has given rise to such calls for transparency.

2.2 Recommendation governance: From the attention economy to attention politics

Given their role in shaping online media consumption, content recommendations from dominant social media platforms exercise an important gatekeeping function with implications for online freedom of expression and media pluralism.[13] Their design, which typically optimises for user engagement, stands accused of surfacing harmful content and distorting online discourse. In response, social media platforms are now increasingly being pressured by policymakers in Europe and elsewhere to curate their recommendations on the basis of various public interest standards, which has in turn raised concerns about the potential for censorship. The following section describes this emergent political economy of recommender governance in greater detail.

As sites of information gatekeeping, recommender systems invite comparison with editorial decisions in the mass media: they both reflect a subjective (and typically commercially motivated) judgement on what content is 'relevant' to their audience.[14]

10 Bernhard Rieder, Ariadne Matamoroz-Fernandez and Oscar Coromina, 'From ranking algorithms to "ranking cultures": Investigating the modulation of visibility in YouTube search results' (2018) 24 *Convergence: The International Journal of Research into New Media Technologies* 50.

11 Rebecca Lewis, 'All of YouTube, Not Just the Algorithm, is a Far-Right Propaganda Machine', *FFWD* (8 January 2020) <https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right- propaganda-machine-29b07b12430> accessed 19 September 2022.

12 Natali Helberger, Katharina Kleinen-Von Königslöw and Rob van der Noll , 'Regulating the new information intermediaries as gatekeepers of information diversity' (2015) 17 *info* 50.

13 On online gatekeeping, see e.g.: Jonathan Zittrain, 'A History of Online Gatekeeping', 19 *Harvard Journal of Law and Technology* 253. Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press 2015). On the specific role of recommendation algorithms in online gatekeeping, see: Helberger, Kleinen-Von Königslöw and Van der Noll, 'Regulating the new information intermediaries' (n 12).

14 Tarleton Gillespie, 'Platforms Intervene' (2015) 1 *Social Media + Society* 1. Van Hoboken, *Search Engine Freedom*. Elizabeth van Couvering (2010) *Search Engine Bias: The Structuration of Traffic on the World Wide Web*. PhD thesis, The London School of Economics and Political Science (LSE).

Where they differ is that content recommendations do not determine *access* to content, like a traditional editor would, but rather *exposure*—a function that Helberger, Kleinen-Von Königslöw and Von der Noll describe as 'indirect editorial influence'.[15] Gillespie makes a similar comparison: 'This may be a gentler intervention than an editor deciding what is a front page story and what isn't worth reporting at all, but it is selection nonetheless, and it matters in many of the same ways'.[16]

Another distinction with traditional media gatekeeping is that platforms tend to process user-generated content, rather than editorially selected content. Even when media organisations use algorithms to personalise content selections, as the New York Times does for instance, they are still drawing from a smaller pool of vetted content than, for instance, YouTube's Recommended Videos. In this regard, Jennifer Cobbe and Jatinder Singh distinguish 'open recommending' of user-generated content by platform services, which is the focus of this paper, from 'curated recommending' of walled garden services such as Netflix, or 'closed recommending' of in-house content by media organisations such as the New York Times.[17] Whilst all these services use complex algorithmic systems to generate personalised content recommendations, the 'open recommending' performed with user-generated content operates at the largest scale and with the greatest diversity of content, serving an essential or even quasi-infrastructural role in many media ecosystems.[18] Given their open nature, they also offer the greatest risk of surfacing harmful or illegal content. In this light, social media recommender systems afford a form of gatekeeping which may at first seem relatively indirect and light-touch, but, given the influential position of a handful of social media platforms, nonetheless has the potential for systemic effects across online media ecosystems.[19]

Social media recommender systems also operate within different organisational and commercial structures than the mass media's editorial selections.[20] Unconstrained by professional and organisational standards of journalism, social media platforms are incentivised to optimise their recommendations primarily for *engagement*.[21] This

15 Helberger, Kleinen-Von Königslöw and Van der Noll, 'Regulating the new information intermediaries' (n 12).

16 Gillespie, 'Platforms intervene' (n 14).

17 Cobbe and Singh, 'Regulating recommending' (n 2).

18 Ben Wagner, 'Free Expression? Dominant information intermediaries as arbiters of internet speech'. In: Martin Moore and Damian Tambini (eds), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (Oxford University Press 2018). José van Dijck, Thomas Poell and Martijn de Waal, *The platform society: Public values in a connective world* (Oxford University Press 2018).

19 Cobbe and Singh, 'Regulating recommending' (n 2).

20 Napoli, 'Social Media and the Public Interest' (n 9).

21 Ibid. Van Dijck, Poell and De Waal, *The Platform Society* (n 18).

'attention economy' logic of recommender systems has drawn extensive criticism from academia, the press, and policymakers, who have highlighted the potential harms that may arise from recommendations optimised for engagement and are increasingly forcing platforms to incorporate alternative design values.

Engagement-optimised social media recommendations are alleged to contribute to a range of harms (though some critiques have stronger empirical grounding than others). To name only a few: content recommenders have been accused of accelerating extremist content and disinformation[22]; polarising audiences and pushing users into homogenous 'filter bubbles' or 'echo chambers'[23]; underserving content on certain social movements and news events, or from particular political viewpoints[24]; exposing children and other vulnerable groups to harmful content[25]; and for reflecting or amplifying societal prejudices and biases against marginalised groups.[26] They have also been accused of intentional political bias and censorship, where platforms allegedly intervened with content on specific issues—though many of these claims remain unverified.[27] Such critiques serve to problematise and politicise the supposed neutrality or objectivity of platform information flows and their determinations of relevance.[28]

Alternative design principles for social media recommendations are now being devised, and, increasingly, implemented in practice. Academics have articulated a range of different values for content recommenders, including 'serendipity'[29],

22 Lewis, 'Alternative Influence' (n 7).

23 Eli Pariser, *The Filter Bubble: What the internet is hiding from you* (Penguin 2011). Axel Bruns, *Are Filter Bubbles Real?* (Polity Press 2019). Frederik Zuiderveen Borgesius and others, 'Should we worry about filter bubbles?' (2016) 5(1) *Internet Policy Review* <https://doi.org/10.14763/2016.1.401> accessed 20 September 2020.

24 Zeynep Tufekci, *Twitter and Tear Gas: The power and fragility of networked protest* (Yale University Press 2017) (describing how the Ferguson protest movement #Blacklivesmatter in the US 'was almost tripped up by Facebook's algorithm').

25 James Bridle, 'Something is wrong on the internet', *Medium* (6 November 2017) <https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2> accessed 25 September 2022. Max Fisher and Amanda Taub, 'On YouTube's Digital Playground, an Open Gate for Pedophiles', *The New York Times* (3 June 2019) <https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html> accessed 19 September 2022.

26 Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press 2018).

27 Thomas Poell and José van Dijck, 'Social Media and New Protest Movements', in: Jean Burgess, Alice Marwick and Thomas Poell (eds.), *The SAGE Handbook of Social Media* (Sage 2017). More broadly, on popular understandings of social media recommendation, see: Motahhare Eslami and others, 'First I "like" it, then I hide it: Folk Theories of Social Feeds' (2016) *CHI '16* 2371.

28 Evgeny Morozov, *To Save Everything, Click Here* (Public Affairs 2014).

29 Natali Helberger, 'Diversity by Design' (2011) 1 *Journal of Information Policy* 441.

'diversity'[30], 'neutrality'[31], 'user choice' and 'user control'[32], and 'agonism'.[33] Each reflects different judgements about the particular risks and opportunities posed by recommender systems, and can be operationalised in countless different ways. But what these proposals have in common, is that they depart from the commercial logics of the attention economy, and instead would have social media recommenders reflect public interests or values.[34] Indeed, several governments across Europe have over the past years proposed to regulate social media recommendations through public law, based on a variety of public interest principles and definitions.[35] And platforms are starting to take note.

Since 2016, major social media platforms claim to have altered their recommender systems in ways that depart from a strictly engagement-driven design, ostensibly in response to concerns over the spread of harmful content. In particular, these changes tend to address content which is not explicitly prohibited by the platform but is nonetheless considered undesirable or unwelcome, such as disinformation and political extremism. Facebook in particular has announced a bevy of such measures. In early 2018, the platform changed their Newsfeed recommendation algorithm to promote content shared by friends and reduce the reach of news pages (presented as a move towards more 'meaningful engagement').[36] In 2019, they announced downranking policies for anti-vaccination content and other 'borderline content' which falls short of violating companies prohibitions.[37] In the same year, they also

30 Natali Helberger, Kari Karppinen and Lucia d'Acunto, 'Exposure diversity as a design principle for recommender systems' (2018) *21 Information, Communication & Society* 191.

31 Frank Pasquale, 'Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power' (2016) 17 *Theoretical Inquiries in Law* 487.

32 Jaron Harambam, Natali Helberger and Joris van Hoboken, 'Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem' (2018) 376 *Philosophical Transactions of the Royal Society* 2133.

33 Kate Crawford, 'Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics' (2016) 41 *Science, Technology, & Human Values* 77.

34 Napoli, 'Social Media and the Public Interest' (n 9). Van Dijck, Poell and De Waal, *The Platform Society* (n 18).

35 See Section 3.2 of this Chapter below.

36 Adam Mosseri, 'Bringing People Closer Together', Facebook Newsroom (11 January 2018). <https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/> accessed 26 September 2022.

37 Monica Bickert, 'Combatting Vaccine Misinformation', *Facebook Newsroom* (7 March 2019) <https://newsroom.fb.com/news/2019/03/combatting-vaccine-misinformation/> accessed 26 September 2022. Facebook, 'How People Help Fight False News, *Facebook Newsroom* (21 June 2018) < https://about.fb.com/news/2018/06/inside-feed-how-people-help-fight-false-news/> accessed 26 September 2022. Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement', *Facebook Notes* (15 November 2018) <https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/> accessed 26 September 2022.

announced a new 'Click-gap' programme to suppress 'low-quality content', which is achieved by analysing the relative popularity of a given item on Facebook compared to its overall web traffic.[38] YouTube also claims to be experimenting intensively with methods to improve recommendation quality and to reduce the spread of harmful and misinforming content. A 2019 blog post claimed that 'in the last year alone, we've made hundreds of changes to improve the quality of recommendations'.[39] Those concerned with combating harmful speech may welcome these interventions, while those concerned with freedom of expression might balk at them. In any case, these examples highlight how recommender systems are increasingly used as a tool for content curation.

A variety of different methods are in play: some interventions target specific speakers or posts, such as Facebook's downranking of false headlines, whereas more fundamental changes to the algorithm have the potential to affect all rankings across the system. Some interventions are decided on a case-by-case basis by human actors, whereas others are automated to a large degree, such as the blacklisting and whitelisting of accounts, keywords or phrases, or analysis of content metadata as in Facebook's aforementioned Click-gap program.[40] In any case, as discussed below, these decisions and their effects are largely opaque to outside stakeholders.

These new forms of curation may be motivated by any number of (perceived) demands or pressures, including political pressures and the threat of government regulation.[41] Social media platforms are embedded in complex governance structures and accountability relationships with a range of different stakeholders: not only governments but also proactive users, civil society actors, and commercial partners may motivate them to intervene in content flows. In any case, it is clear that their actions cannot be explained solely through 'attention economy' accounts about engagement optimisation. This is not to deny the commercial, profit-seeking nature

38 Guy Rosen, 'Remove, Reduce, Inform: New Steps to Manage Problematic Content', *Facebook Newsroom* (10 April 2019) <https://newsroom.fb.com/news/2019/04/remove-reduce-inform-new-steps/> accessed 9 November 2022.

39 YouTube, 'Continuing our work to improve recommendations on YouTube', *YouTube Official Blog* (25 January 2019) <https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html> accessed 26 September 2022.

40 On algorithmic blacklisting and whitelisting, see: Jeff Gary and Ashkan Soltani, 'First Things First: Online Advertising Practices and Their Effects on Platform Speech', *Knight First Amendment Institute* (21 August 2019) <https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech> accessed 9 November 2022.

41 e.g. Damian Tambini, Danilo Leonardi and Christopher Marsden, 'The privatisation of censorship: self-regulation and freedom of expression', in Damian Tambini, Danilo Leonardi and Chris Marsden, *Codifying cyberspace: communications self-regulation in the age of internet convergence* (Routledge 2008).

of social media platforms, but simply to recognise that their economic self-interest may require them to take into account social and political conditions. In other words: recommendation gatekeeping is not simply a matter of attention economy, but also, and increasingly, of attention politics.

This struggle over the future of recommendation gatekeeping does not appear to have definitive answers or solutions. Public interest concepts such as media pluralism—i.e. the appropriate structure or balance of available media in a given polity—cannot, as Kari Karppinen observers, be 'solved' objectively or definitively.[42] Indeed, it is worth noting the tensions between current design proposals, such as 'diversity' and 'trustworthiness' on the one hand and 'non-discrimination' and 'neutrality' on the other; whereas the former group would require recommenders to seek out and prioritise certain content, the latter could arguably prohibit such differentiation. We need not expect a consensus on such issues to emerge soon: the recognition is growing that there is no such thing as a neutral recommender system, and what remains is a fundamentally political and value-laden question as to what types of content should be prioritised across different segments of the population. It will likely continue to be contested for the foreseeable future, as a new frontier in media governance.[43]

2.3 'Obscured obscuring': The opacity of social media recommendations

A commonly criticised aspect of recommendation governance is that it is deeply opaque. While it is clear *that* platforms increasingly curate their recommendations for various forms of content regulation, *how* they do so is difficult to observe and understand. Recommender systems are perceived as 'black boxes', whose internal logics are inscrutable and their outputs unpredictable, creating a barrier to holding these systems accountable.[44] Gillespie memorably warns against 'the *obscured obscuring* of contentious material from the public sphere' (emphasis added), which 'raises a new challenge to the dynamics of public contestation and free speech'.[45] So what makes these systems so opaque? Their lack of transparency is multifaceted, and results from both technical and legal factors.

Taken in its most basic sense, transparency can be said to refer to 'the disclosure of

42 Kari Karppinen, *Rethinking Media Pluralism* (Fordham University Press 2013).

43 Ibid. Natali Helberger, 'On the democratic role of news recommenders' (2019) 7 *Digital Journalism* 993.

44 Pasquale, *The Black Box Society* (n 3).

45 Tarleton Gillespie, *Custodians of the internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018).

certain information that may not previously have been visible or publicly available'.[46] In the context of recommender systems, concerns over transparency often refer to the specific algorithms used to produce recommendations. But other aspects of recommender systems are also opaque, such as the outputs (what recommendations are made?) and inputs (user content & metadata, behavioural data, etc.) In addition, transparency can also refer to the human agents and organisational structures involved in designing and operating this system. At present, many influential content recommenders lack transparency on each of these issues—from the algorithm as such to its inputs and outputs and the surrounding institutions.

To start with the recommendation algorithms: these are obscure due to their technical complexity as well as intentional corporate secrecy.[47] Given their scale and complexity, these algorithms are often ill-suited to 'human scale comprehension', and it is difficult even for experts to develop concrete, causal explanations for specific outcomes.[48] Some platforms now offer individualised 'explanation' features, such as Facebook's Why Am I Seeing This? feature, but such efforts have been criticised for failing to meaningfully describe the full complexity of the algorithm's operations.[49] Platforms could in theory publish their algorithms in full and enable outside study, but they have reasons to keep them confidential. First, platforms commonly argue that recommender system design involves commercially valuable trade secrets.[50] Second, confidentiality of the algorithm may in some cases be necessary to prevent users from 'gaming' the system and undermining its gatekeeping function.[51] For instance, if platforms were to publish their keyword blacklists, this could help sophisticated spammers to avoid being downranked in this way. Third, documentation of recommender systems algorithms could in some cases jeopardise the privacy of platform users, if this algorithm was developed on the basis of user profile data.

46 Oana Albu and Mikkel Flyverblom, 'Organizational Transparency: Conceptualizations, Conditions, and Consequences' (2016) 58 *Business & Society* 268. Robert Gorwa and Timothy Garton Ash, 'Democratic Transparency in the Platform Society' in: Nate Persily and Joshua Tucker (eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge University Press 2020).

47 Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2017) 3(1) *Big Data & Society* <https://doi.org/10.1177/2053951715622512> accessed 15 September 2022.

48 Ibid.

49 Lilian Edwards and Michael Veale, 'Slave to the algorithm? Why a "Right to an Explanation" is probably not the remedy you are looking for' (2017) 16 *Duke Law & Technology Review* 18.

50 Burrell, 'How the machine 'thinks' (n 47).

51 Pasquale argues that platforms' claims in this context should not be taken at face value; only under certain conditions will transparency enable 'gaming', and platforms' claims may in fact be motivated by proprietary concerns. Pasquale, *The Black Box Society* (n 3). See also: Nicholas Diakopoulos, 'Accountability in Algorithmic Decision Making' (2016) 59 *Communications of the ACM* 56.

But a clear view of inputs and outputs is also crucial to understanding recommender systems. As discussed in Section 2.1, the sociotechnical perspective on recommender systems highlights that the significance of algorithms is very much contextual, as outputs are co-determined by user behaviour. To understand their functioning in practice, a view on their outputs is therefore necessary. Rieder, Matamoroz-Fernandez and Coromina conclude that for the opacity of recommender systems, 'access to the mythical source code would not solve this problem.'[52] Instead, they argue for research methods focused firstly on the *outcomes* of these systems, in terms of what recommendation patterns are generated on particular issues and for particular publics, and how these change over time.[53]

But the study of recommender system outputs and outcomes is restricted in several ways, first and foremost as a result of their personalisation. Since each user is served a personalised selection of recommendations, it is difficult for any individual observer to make generalisable conclusions about the outputs of the system as a whole.[54] All we know is our own news feeds; as to what others are seeing, we can only guess. Researchers have attempted to counteract this obscuring effect of personalisation through survey techniques, which mobilise a large number of accounts (either bots or human volunteers) to assemble data about the platform's outputs.[55] One notable example out of many is the German *Datenspende* project, which tracked Google search results during a 2017 election for over 4000 participants.[56] However, even the most ambitious and elaborate of these methods can only provide snapshots, and do not come close to a comprehensive or systemic view of platform traffic flows. Worse still, platforms can and have restricted these processes contractually by way of their Terms of Service, and technically by way of blocking such tools. For instance, Facebook recently blocked a popular data scraping tool by ProPublica, citing violations of its Terms of Service.[57]

52 Rieder, Matamoroz-Fernandez and Coromina, 'From ranking algorithms to "ranking cultures"' (n 10).

53 Ibid.

54 Balazs Bodó and others, 'Tackling the Algorithmic Control Crisis: The Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents' (2018) 19 *Yale Journal of Law and Technology* 133.

55 Eduardo Hargreaves and others, 'Biases in the Facebook News Feed: a Case Study on the Italian Elections' (2018) *International Symposium on Foundations of Open Source Intelligence and Security Informatics, In conjunction with IEEE/ACM ASONAM* <https://hal.inria.fr/hal-01907069> accessed 19 September 2022. Christian Sandvig and others, 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms' (2014), Paper presented to *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* <https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf> accessed 26 September 2022.

56 Tobias Krafft, Michael Gamer and Katharina Zweig, 'Wer sieht was? Personalisierung, Regionalisierung un die Frage nach der Filterblase in Googles Suchmacschine' (Research Report Project #Datenspende 2018) <https://www.blm.de/files/pdf2/bericht-datenspende---wer-sieht-was-auf-google.pdf> accessed 19 September 2022.

57 Jeremy Merrill and Ariana Tobin, 'Facebook Moves to Block Ad Transparency Tools —Including Ours', *ProPublica* (28 January 2019) <https://www.propublica.org/article/facebook-blocks-ad-transparency-tools> accessed 19 September 2022.

Besides independent surveying, one of the most important sources of data regarding recommender systems has been their public APIs, through which outside researchers can download platform data in bulk. But these have come under significant pressure over the past years. Since the Cambridge Analytica scandal, in which academics helped to leak and abuse large sets of user data from Facebook, important APIs have incurred major restrictions in their functionality. This development, which Axel Bruns describes as the 'APIcalypse', has been the demise of many widely-used research tools and methodologies, both commercial and academic.[58] Of course, the quality of API access differs between platforms; for instance, YouTube and Twitter offer relatively generous public research APIs, whereas Instagram's was recently shut down entirely.[59] Regardless of what information is currently available, Deen Freelon warns that, since platforms have no binding obligation to maintain these systems in any consistent manner, the situation is fundamentally precarious: 'we find ourselves in a situation where heavy investment in teaching and learning platform-specific methods can be rendered useless overnight'.[60]

Through code and through contract, then, platforms are able to obstruct independent study of their recommender systems, leaving even the basic outputs unclear.[61] In this sense, most platform content recommenders are even less transparent than the prototypical 'black box': not only is it unclear *why* certain decisions are being made, it is simply unclear *what* decisions are being made in the first place. This is an important contrast with other prominent debates in algorithmic governance, such as, for instance, judicial sentencing algorithms, where the algorithm may be secret but the ultimate decisions are still a matter of public record.[62] It is also a noteworthy contrast with mass media content distribution of press, radio and television, where outputs are equally a matter of public record and thus render the editorial line of a given outlet readily identifiable for any and all audience members.[63] The personalised gatekeeping

58 Axel Bruns, 'After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research' (2019) 22 *Information, Communication & Society* 1544.

59 Munger and Philips, 'Right-Wing YouTube' (n 8).

60 Deen Freelon, 'Computational research in the post-API age' (2018) 35 *Political Communication* 4.

61 Julie Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press 2019).

62 On transparency and accountability in judicial sentencing algorithm, see e.g.: Julia Angwin, 'Make Algorithms Accountable', *The New York Times* (1 August 2016). <https://www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html> accessed 16 September 2022. Alyssa M. Carlson, 'The Need for Transparency in the Age of Predictive Sentencing Algorithms' (2017) 103 *Iowa Law Review* 303.

63 This comparison has been made in numerous commentaries on news personalisation, e.g.: Bodó and others, 'Tackling the algorithmic control crisis' (n 54). Yochai Benkler, Robert Faris and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press 2018). It is worth noting that personalisation is also a growing trend in (the online editions of) mass media organizations, posing similar problems for the study of these outlets: Nick Diakopoulos and Michael Koliska, 'Algorithmic Transparency in the News Media' (2016) 5 *Digital Journalism* 809.

of recommender systems, by contrast, is difficult for any outsider to observe at a systemic level.

A final aspect of opacity relates to the organisations surrounding social media recommender systems, which also tend to be poorly documented. As mentioned, speculation abounds regarding the possibility of human interventions in important content recommender systems, such as YouTube's Trending Videos and Facebook's Newsfeed, but there are few conclusive or authoritative sources of information about these platforms' internal operations. In their absence, conjectural 'folk theories' and 'algorithmic lore' proliferate.[64] Platforms occasionally disclose specific policies, such as Facebook's aforementioned fact-checking partnerships: this program's policies are outlined on the Facebook website, and fact-checkers are required to publish explanations for each fact-checking decision on their respective websites, known as 'reference articles'. Unfortunately, there is no central repository of these reference articles, leaving them scattered across dozens of websites without any clear standardisation or comprehensive overview. More fundamentally, these partnered fact-checks are but one example out of many possible downranking interventions, which may not be subject to any clear transparency policies at all. How else are platforms and their affiliates intervening? In the most extreme cases, recommendations may not be automated at all, but instead be curated entirely by human operators. Facebook's by-now notorious Trending Topics was discovered in 2016 to be manually curated by Facebook staff, rather than by an automated algorithmic process, and these revelations quickly prompted accusations of political bias.[65] As such stories illustrate, the opacity of social media recommendations relates not only to their technical specifications, but also the organisational structures in which they are embedded. In the words of Ananny and Crawford, transparency in algorithmic systems should take into account 'not just code and data but an *assemblage* of human and non-human actors.'[66] Indeed, as platforms are being pushed to take more proactive and substantive responsibility for recommendation outcomes, corrective interventions in recommender systems are likely to expand in future.

All this means that the quasi-editorial influence exercised by platform recommendations is difficult for outside stakeholders to study, much less evaluate

64 Sophie Bishop, 'Algorithmic Experts: Selling Algorithmic Lore on YouTube' (2020) 6 *New Media + Society* 1.

65 Michael Nunez, 'Former Facebook Workers: We Routinely Suppressed Conservative News', *Gizmodo* (5 September 2016) <https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006> accessed 19 September 2022.

66 Ananny and Crawford, 'Seeing without knowing' (n 4). Margot Kaminski, 'Understanding Transparency in Algorithmic Accountability' in: Woodrow Barfield (ed.), *Cambridge Handbook of the Law of Algorithms* (Cambridge University Press 2020).

or hold accountable. How platforms design and adapt their algorithmic systems is effectively hidden from public knowledge, as are the actual outputs and outcomes of their choices.

Though platforms have devised a number of self-regulatory transparency measures, these have broadly failed to assuage criticisms. Relevant efforts include user-facing notices (e.g. Facebook's 'Why Am I Seeing This' feature discussed above) as well as data sharing projects with civil society. They tend to be met with scepticism for several reasons. Firstly, creating meaningful transparency arguably runs counter to platforms' incentives: they have a commercial interest in monetising traffic data and insights, and thus in keeping this information exclusive, as well as a political interest in avoiding negative publicity.[67] Indeed, Facebook's flagship Social Science One Initiative has been marred with delays and controversies; whilst Facebook cites legal concerns over data protection compliance, others blame a lack of incentives and political will.[68] Even the European Data Protection Supervisor recently argued as much: 'It would appear therefore that the reluctance to give access to genuine researchers is motivated not so much by data protection concerns as by the absence of business incentive to invest effort in disclosing or being transparent about the volume and nature of data they control.'[69] Such considerations may explain the recent attention for government regulation of transparency.

3. State of play: Regulating recommendation transparency in Europe

The law and policy literature displays a strong consensus around the need for greater transparency in social media governance, particularly as regard content recommender systems.[70] Yet it is also widely acknowledged that 'transparency' is an ambiguous

67 Bruns, 'After the APICalypse' (n 58).

68 European Data Protection Supervisor, 'A Preliminary Opinion on data protection and scientific research' (2020) <https://edps.europa.eu/sites/ed72p/files/publication/20-01- 06_opinion_research_en.pdf> accessed 27 September 2022.

69 Ibid.

70 e.g. Cobbe and Singh, 'Regulating Recommending' (n 2). Gillespie, *Custodians of the internet* (n 45). Van Dijck, Poell and De Waal, *The Platform Society* (n 18). Pasquale, *The Black Box Society* (n 3). Gorwa and Garton Ash, 'Democratic Transparency in the Platform Society' (n 72). Daphne Keller and Paddy Leerssen, 'Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation', in: Persily N. and Tucker J (eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge University Press 2020). Ben Wagner and others, 'Regulating transparency?: Facebook, Twitter and the German Network Enforcement Act' (2020) *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 261. Nick Suzor, 'Digital Constitutionalism': Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4(3) *Social Media + Society* <https://doi.org/10.1177/205630511878781> accessed 19 September 2022. Laidlaw, *Regulating Speech in Cyberspace* (n 13).

concept that can be operationalised in numerous ways, particularly as regards such complex technological phenomena as recommender systems. As Robert Gorwa and Timothy Garton Ash argue, 'transparency in practice is deeply political, contested, and oftentimes problematic'—or, more bluntly by Amitai Etzioni, 'a form of regulation by other means'.[71] The following section reviews European plans to regulate transparency in social media recommender systems, and the types of accountability they pursue.

Transparency measures can be analysed in numerous ways. A substantial literature of different transparency taxonomies has emerged, with early work focusing on government transparency but later turning to address private actors and, more recently, platforms and algorithmic systems in particular.[72] Transparency has accordingly been conceptualised in terms of its subjects, formats, rationales, timing, effectiveness, and so many other factors.[73] A recurring theme in this literature, which will guide the discussion in this paper, is the question *to whom* transparency is offered. This accords with the common understanding of transparency and accountability as relational concepts, which are defined by the stakeholders they serve.[74] Following Ananny and Crawford, interrogating the relationship between a proposed transparency measure and its intended accountability outcome, must start with the question *to whom* accountability will be rendered.[75] A similar relational focus can also be seen in, for instance, the work of David Weil, Mary Graham and Archon Fung on 'targeted transparency'.[76] By examining the audiences that transparency measures serve, we can begin to chart the more fundamental visions of platform accountability that inform these measures.

In the past years, European policymakers have undertaken several different initiatives to regulate the transparency of social media recommendations. This by now complex

71 Amitai Etzioni, 'Is Transparency the Best Disinfectant?' (2010) 18 *The Journal of Political Philosophy* 389. Gorwa and Garton Ash, 'Democratic Transparency in the Platform Society' (n 72).

72 Gorwa and Garton Ash, 'Democratic Transparency in the Platform Society' (n 72). Ananny and Crawford, 'Seeing without knowing' (n 4). Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 66).

73 e.g. David Heald, 'Varieties of transparency' in David Heald and Christopher Hood and David Heald (eds.). In *Transparency: The key to better governance?* (Oxford University Press for the British Academy 2006).

74 E.g. Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 *European Law Journal* 447. Richard Mulgan, 'Accountability: An Ever-Expanding Concept?' (2000) 78 *Public Administration* 555.

75 Ananny and Crawford, 'Seeing without knowing' (n 4), citing Theodore Glasser, 'Three views on accountability', in Everette Dennis, Donald Gillmor and Theodore Glasser (eds.), *Media Freedom and Accountability* (Praeger 1989).

76 Archon Fung, David Weil and Mary Graham, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge University Press 2007).

and fragmented landscape includes horizontal instruments, such as competition law and data protection law, which are not tailored to social media governance in particular but may still have some spillover benefits for its purposes. More recently we also see the emergence of several sectoral proposals that lay out a specific vision on transparency for social media recommendations in particular. Most of the latter instruments are rooted in media pluralism policy, but they also target other public interest considerations such as the combating of online disinformation. Despite the variety of rules in play, the transparency measures contained in these instruments can be grouped into three general categories, aimed at three different sets of stakeholders: (1) user-facing disclosures, which aim to channel information towards individual users in order to empower them in relation to the content recommender system, (2) government oversight, which appoints a public entity to monitor recommender systems for compliance with publicly-regulated standards, and (3) partnerships with academia and civil society, which enable these stakeholders to research and critique recommender systems. Each of these is discussed below.

	Disclosures for users (end-users and content providers)	**Disclosures for public authorities**	**Disclosures for academia and civil society**
Disclosure Form	Disclaimers, notices, 'explanations'	Audits, reporting requirements	Data-sharing partnerships, initiatives, observatories, etc.
Associated Accountability Standard(s)	User choice / revealed preference	Public standard setting (e.g. non-discrimination, pluralism, trustworthiness)	Various/undefined.

Table 1: Typology of disclosure rules for social media recommenders in Europe

At the outset, it should be noted that these different types of transparency are by no means mutually exclusive; rather, they reflect the growing consensus that platform governance requires a multistakeholder approach.[77] Accordingly, most scholars defend a variegated or tiered approach to transparency and accountability in this space, such as Frank Pasquale's 'qualified transparency' model and Andrew Tutt's 'Spectrum of Disclosure'.[78] As the following section shows, European policy is

77 e.g. Laidlaw, *Regulating Speech in Cyberspace* (n 13). Cobbe and Singh, 'Regulating recommending' (n 2). Chris Marsden, Trisha Meyer and Ian Brown, 'Platform values and democratic elections: How can the law regulate digital disinformation?' (2020) 36 *Computer Law & Security Review* 105373.

78 Pasquale, *The Black Box Society* (n 3). Andrew Tutt, 'An FDA for Algorithms' (2017) 69 *Administrative Law Review* 83. Tal Zarsky, 'Transparent Predictions' (2013) 4 *University of Illinois Law Review* 1503. Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 66).

developing such a tiered approach, in which understandable, simplified information is channelled towards individual end-users, and detailed, sensitive information is shared confidentially with experts in government and civil society.[79]

What appears to distinguish social media from other areas of platform governance, is the growing emphasis on transparency for civil society and academia, engaging what Archon Fung describes as 'the civic immune system' and Mark Bovens as social and political accountability (as distinct from legal or administrative accountability).[80] Including these actors may seem relatively uncontroversial, relative to direct command-and-control regulation, and indeed it appears that their inclusion is motivated by the politically sensitive nature of media governance. But it is in defining and institutionalising these notionally independent groups that problems are likely to emerge. Maintaining the inclusiveness and independence of such efforts, and ultimately their legitimacy, necessitates that policymakers should also turn their attention towards developing a robust vision for *public* data access for recommender systems, without restrictions on who can access the data involved.

3.1 User-facing disclaimers

Perhaps the most common approach to regulating transparency in recommender systems is to require disclosures for individual users. The aim of transparency in this context is to *inform* users about their available options so as to help them realise their own preferences, appealing to such values as individual autonomy, agency and trust.[81] If platforms fail to do so, users can, in theory, respond by exiting the platform and taking their activity elsewhere. Napoli describes this as the 'individualist model' of social media governance, in which platforms are required to 'provide an enabling environment in which individual responsibility and autonomy can be realised in relation to the production, dissemination, and consumption of news and information'.[82] It should be noted that the category of 'users' in the context of social media platforms includes not only the consumers of content, but also the providers of content, ranging from amateur vloggers to professional influencers and major media organisations. With that in mind, user-facing transparency can also appeal to principles of competition, fairness and diversity in online media markets.

79 Pasquale, *The Black Box Society* (n 3).

80 Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 *European Law Journal* 447.

81 Max van Drunen, Natali Helberger and Mariella Bastian, 'Know your algorithm: what media organizations need to explain to their users about news personalization' (2019) 9 *International Data Privacy Law* 220.

82 Napoli, 'Social Media and The Public Interest' (n 9).

This user-facing approach to transparency can be seen in several European instruments. The General Data Protection Regulation (GDPR) grants platform users a bundle of individual rights. Article 5 lists 'transparency' as one of the Regulation's key principles, and users are granted a bevy of information and notice rights about personal data processing under Article 12-14. More specifically, under Article 22, data subjects may under certain circumstances have the right to opt out of such automated decisions, and also enjoy a bundle of information rights collectively known as the 'right to an explanation'.[83]

Given that the GDPR focuses on data protection, rather than media governance or platform gatekeeping per se, the information acquired in this way could be of only tangential relevance to the study of platform gatekeeping. However, expansive interpretations may be possible: Max van Drunen, Natali Helberger and Mariella Bastian have studied this right as it applies to news recommender systems, and conclude that these provisions should be interpreted *contextually* as a means to empower data subjects in their capacity as news consumers.[84] On this basis, they argue that users of recommender systems are entitled to a range of information about e.g. the parties able to influence editorial decisions, the profiles that the algorithms construct about them, and the algorithm's metrics and factors.[85] In such a reading, the GDPR could in theory be a source of insights about platform gatekeeping decisions. It remains to be seen whether such access rights will find much usage with the average end user and in fact serve as a source of empowerment in practice.

Another relevant horizontal instrument is the Regulation on Promoting Fairness and Transparency for Business Users of Online Intermediation Services (Platform-to-Business Regulation).[86] This Regulation affects a different category of users: not *consumers* of social media content, but rather *producers*, who are granted certain notice rights in relation to recommender systems under Article 5. This provision requires platforms to disclose, *inter alia*, 'the characteristics of the goods and services offered to consumers through the

83 Andrew Selbst and Julia Powles, 'Meaningful information and the right to explanation' (2017) 7 *International Data Privacy Law* 233. Margot Kaminski, 'The Right to Explanation, Explained' (2019) 34 *Berkeley Technology Law Journal* 189.

84 Van Drunen, Helberger and Bastian, 'Know your algorithm' (n 81).

85 Van Drunen, Helberger and Bastian (n 81) also recommend disclosures about e.g. sources (the origin of the information). However, these arguments appear to focus on recommender systems operated by *news organisations*, which have direct knowledge and control of such elements. Such disclosures do not appear to be as feasible in the context of platforms.

86 Regulation 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services ('P2B Regulation').

online intermediation services or the online search engine'.[87] For sophisticated content providers who rely on social media, such as newspapers and other media outlets, this could be an additional way to adapt to changes in recommendation algorithms, and potentially to detect unlawful or abusive forms of discrimination.[88]

New proposals particular to media governance are also emerging. The *Medienstaatsvertrag*, proposed in 2018 by the German broadcast authority, requires media intermediaries to disclose the selection criteria that determine the sorting and presentation of content. These include 'the central criteria of aggregation, selection and presentation of content and their weight, including information about the function of the algorithms used'.[89] Addressed towards end-users, they must be made in 'understandable language', and in 'in easily recognisable, directly accessible and constantly available formats'.[90] Comparable recommendations are made in the EU Code of Practice on Disinformation, which is a co-regulatory instrument signed by Facebook, Google and Twitter under the guidance of the European Commission.[91] These companies must 'consider empowering users with tools enabling a customised and interactive online experience so as to facilitate content discovery and access to different news sources representing alternative viewpoints, also providing them with easily-accessible tools to report Disinformation.'[92] A number of Council of Europe recommendations also emphasises the importance of informing and empowering users. For instance, Recommendation 2018-1 on media pluralism and transparency of media ownership calls on states to encourage platforms to 'provide clear information to users on how to find, access and derive maximum benefit from the wide range of content that is available'.[93]

---

87 P2B Regulation, Article 5.

88 Alessio Cornia and others, *Private Sector News, Social Media Distribution ,and Algorithm Change* (Research Report Reuters Institute Digital News Project Report 2018) <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-10/Cornia_Private_Sector_News_FINAL.pdf> accessed 25 September 2020.

89 *Staatsvertrag zur Modernisierung der Medienordnung in Deutschland – Entwurf* ('*Mediensta-atsvertrag*'), Art 53(d)(translation mine).

90 Ibid.

91 EU Code of Practice on Disinformation (European Commission 2018) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> accessed 26 September 2022. Although this document has been characterised as self-regulatory, it is better understood as co-regulatory given the European Commission's close involvement in its development and implementation. See: Aleksandra Kuczerawy, 'Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?', in: Elzbieta Kużelewska and others (eds.), *Disinformation and Digital Media as a Challenge for Democracy* (Intersentia 2021).

92 EU Code of Practice on Disinformation, p. 3.

93 Council of Europe, 'Recommendation of the Committee of Ministers to member States on media pluralism and transparency of media ownership: Guidelines on media pluralism and transparency of media ownership' (2018) CM/Rec(2018)1. For a critical discussion of this user-centric strand in Council of Europe standards, see: Napoli, 'Social media and the public interest' (n 9).

Evidently, the notion that individual user rights should 'empower' users vis-à-vis social media recommender systems is widespread in European policy circles. But there are also important limitations to these user-centric approaches, both practical and principled. As a practical matter, informing users about complex systems such as content recommenders is difficult, and not straightforwardly achieved through disclaimers or notices. As stated, the complexity of recommender systems renders 'algorithmic explanations' difficult if not in possible, certainly in formats that are digestible to the average end-user. Evidence from privacy and consumer protection law scholarship shows that user-facing notices on social media platforms and other websites are routinely neglected by the vast majority of users.[94] And even where information is made to be 'simplified' and 'understandable', as media governance instruments are now requiring, these effects are likely to persist—the most infamous precedent being the cookie consent notices required under EU privacy law.[95]

Even if fully informed, individual users may simply lack the market *power* to depart from dominant platform offerings. Due to such well-documented dynamics as market concentration, network effects, and user-lock in, it may be costly or even impossible for users to switch to viable alternative platforms.[96] In this sense, transparency towards users may not have full effect if it is 'disconnected from power' to actually change outcomes.[97]

Given these manifold constraints on user-facing disclosures, it remains debatable whether expanding individual transparency rights will have much impact on the average user. A greater impact might be expected with more sophisticated platform users, such as professional content providers or media organisations who rely on social media to ply their trade. Also worth noting is that academics and journalists are starting to experiment with access rights under the GDPR (exercised directly by the researcher or by indirectly with the help of volunteers) as a source of data; as Jef Ausloos argues, individualised user rights may thus have unexpected spillovers from their stated goal of individual empowerment to a more collective and social forms of accountability pursued by academics and civil society actors.[98]

94 Frederik Zuiderveen Borgesius, 'Behavioural sciences and the regulation of privacy on the internet'. In Anne-Lise Sibony and Alberto Alemanno (eds.), *Nudging and the law: what can EU law learn from behavioural sciences?* (Hart Publishing 2015).

95 Ibid.

96 Patrick Barwise and Leo Watkins, 'The evolution of digital dominance: how and why we got to GAFA', in: Martin Moore and Damian Tambini (eds.), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (Oxford University Press 2018).

97 Ananny and Crawford, 'Seeing without knowing' (n 4).

98 Jef Ausloos, 'GDPR Transparency as a Research Method' (2019) SSRN Working paper. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3465680> accessed 16 September 2022.

More fundamentally, the ideal of 'user empowerment' can be criticised as overly individualistic, and endorsing a 'neoliberal model of agency'.[99] While informing users may serve to enhance choice and competition, as Napoli points, media governance has typically not allowed the public interest to be defined exclusively by these market-based ordering principles. The view that 'the public's interest, then, defines the public interest'[100] is marginal, certainly in the European tradition. Rather, media policy has also relied on public and collective forms of governance, including government oversight and professional self-regulation, in order to safeguard public values that risk being underserved in a laissez-faire environment, such as pluralism, diversity, child protection, and localism.[101] Of course, individualist values such as choice, autonomy, competition and agency may still be recognised as important within a broader conception of the public interest. But to *equate* them with the public interest, is to oversimplify the challenges of media governance.

3.2 Government oversight

Several European institutions have proposed government oversight of social media recommendations, in order to safeguards public interest principles such as diversity or child protection,[102] enforced by independent regulatory agencies. In terms of transparency, this form of governance relies on reporting duties for platforms and/or auditing powers vested in the regulator.[103] With relevant expertise and the ability to ensure confidentiality of information disclosed, governments can process more detailed information than user-facing notices allow. Government oversight frameworks for social media recommenders are not yet as commonplace in Europe as user-facing disclaimers, but a number of horizontal instruments apply, and several sectoral proposals have surfaced in recent years.

One of the most advanced proposals for public oversight of social media recommendations is Germany's aforementioned *Medienstaatsvertrag*. Its key requirement would be non-discrimination. Under this framework, social media platforms 'may not unfairly disadvantage (directly or indirectly) or treat differently providers of journalistic editorial content to the extent that the intermediary has

99 Ananny and Crawford, 'Seeing without knowing' (n 4). Elettra Bietti, 'Consent as a Free Pass: Platform Power and the Limits of the Informational Turn' (2020) 40 *Pace Law Review* 307.

100 Mark Fowler and Daniel Brenner, 'A Market Place Approach to Broadcast Regulation' (1982) 60 *Texas Law Review* 207.

101 Napoli, 'Social Media and the Public Interest' (n 9).

102 Ibid.

103 Caveat: as Kaminski observes, governments commonly rely on knowledge sourced from other societal actors. Accordingly, transparency measures aimed at other stakeholders may also offer indirect benefits to government oversight. Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 66).

potentially a significant influence on their visibility'.[104] German broadcast regulators at the federal and local level would be empowered to set detailed standards for social media recommender design, and to request documentation from platforms about their activities.[105] In the Netherlands, the Dutch State Commission on the Parliamentary System has proposed a comparable 'independent entity' to monitor social media recommenders, but in contrast to the German proposal their mandate would focus not on non-discrimination but rather on maintaining 'diversity' and avoiding 'bias'.[106] 'If a strong bias can be observed which does not correspond to the information offered by the users themselves on the platform, or if that bias suddenly changes during an election period, this entity can remark on this and ask the company for a response.'[107]

At EU level, the main instrument for media regulation is the Audiovisual Media Services Directive. However, it does not contain any particular rules related to recommender systems. More relevant for our purposes is the EU Code of Practice on Disinformation, which requires signatories to '[d]ilute the visibility of disinformation by improving the findability of trustworthy content' and to 'invest in technological means to prioritise relevant, authentic, and authoritative information where appropriate in search, feeds, or other automatically ranked distribution channels'.[108] However, this Code is a non-binding co-regulatory instrument, and it lacks any concrete sanctions or enforcement mechanisms; platforms were merely expected to self-report their compliance efforts in the months prior to the European Election of May 2019. In terms of transparency, then, it is not armed with the same investigative powers as a conventional regulatory agency. Binding regulation at EU level does appear to be under consideration: leaked policy briefs from the Von der Leyen Commission from 2019 envisages 'a dedicated regulatory structure' for the oversight of online platforms, with a particular focus on creating transparency.[109]

The Council of Europe has also developed standards on the need for government oversight of content recommenders, emphasising diversity or pluralism as a guiding principle. Their Committee of Ministers has recommended that '[s]tates should encourage social

104 Medienstaatsvertrag (proposed), Article 53(e) (translation mine)
105 Medienstaatsvertrag (proposed) , Article 53(f) (translation mine)
106 Johan Remkes and others, Lage Drempels, Hoge Dijken: eindrapport (Staatscommissie Parlementair Stelsel 2018) <staatscommissieparlementairstelsel.nl/documenten/rapporten/samenvattingen/12/13/eindrapport> accessed 26 September 2022.
107 Ibid.
108 EU Code of Practice on Disinformation, p. 3.
109 Alexander Fanta and Thomas Rudl, 'Leaked document: EU Commission mulls new law to regulate online platforms', *Netzpolitik* (16 July 2019). <https://netzpolitik.org/2019/leaked-document-eu-commission-mulls-new-law-to-regulate-online-platforms/#spendenleiste> accessed 15 September 2022.

media, media, search and recommendation engines and other intermediaries which use algorithms … to engage in open, independent, transparent and participatory initiatives that seek to improve these distribution processes in order to enhance users' effective exposure to the broadest possible diversity of media content.'[110] In contrast to the foregoing examples, this wording does not expressly refer to regulatory agencies but instead describes in more general terms a need for 'open' and 'participatory' institutions or initiatives, suggesting a more co-regulatory or multistakeholder approach. The state's more modest role lies in 'encouraging' such efforts.

Government oversight of platform recommender systems can also be found in horizontal instruments in data protection and competition law. The General Data Protection Regulation sets limits and conditions on the processing of personal data by content recommender systems, which constrains their ability to personalise content. These rules can be enforced privately by data subjects, but also by national data protection authorities (DPAs). Likewise, competition law constrains dominant platforms in their ability to discriminate between commercial actors on their platform, as a potential abuse of their dominant position.[111] This standard is most relevant for vertically integrated platforms, which also produce their own content and thus have an incentive to discriminate against rival content providers.[112] Both data protection and competition authorities are vested with a bevy of investigative powers, such as requesting documentation and performing audits. These frameworks do not directly address the same public interest concerns as media policy, so it is unlikely that these efforts will be targeted directly at studying media governance issues such as pluralism or disinformation. Nonetheless, their research may still have spillover effects between regulatory fields, potentially revealing information that is relevant to media governance.[113]

Government oversight of social media recommendations faces many significant challenges, both practical and principled. Most straightforward is the fact that government authorities are capacity-constrained, particularly as regards the technical expertise required to perform complex algorithmic auditing, and in relation to the sheer scale and scope of potential research issues at stake in social media governance. This is especially true for horizontal agencies such as competition and data protection

110 Council of Europe, CM/Rec(2018)1, Article 2(5).

111 Treaty on the Functioning of the European Union, Article 102.

112 Helberger, Kleinen-Von Konigslow, von der Noll, 'Regulating the new information intermediaries as gatekeepers of information diversity' (n 12).

113 European Data Protection Supervisor, 'EDPS Opinion 8/2016 on coherent enforcement of fundamental rights in the age of big data' (2016) <https://edps.europa.eu/sites/edp/files/publication/16-09-23_bigdata_opinion_en.pdf> accessed 26 September 2022.

authorities, for whom social media recommender systems risk being overshadowed and overlooked in an extensive, economy-wide portfolio. Sectoral proposals, on the other hand, would in many cases require the creation of *entirely new* oversight bodies, or for traditional broadcast regulators to develop radically new forms of expertise. What makes this particularly challenging is that, in Europe, media policy is largely a national affair, without a clear institution at EU level capable of performing a monitoring role. Indeed, EU governments have repeatedly shot down proposals for creating a supranational media authority.[114] National-level action in this space, on the other hand, could result in a duplication and fragmentation of efforts.

It is worth nothing that, given these capacity constraints on government monitoring, government agencies commonly rely on knowledge sourced from other societal actors, through such formats as public consultations, expert hearings, and complaint procedures. Therefore, as Margot Kaminski observes, transparency measures aimed at third parties such as users, civil society and other stakeholders can also serve indirectly to enhance accountability to public regulation.[115]

Principled objections to government monitoring as a form of transparency are also possible. As discussed, public standard setting for recommender systems necessarily involves (quasi-)editorial judgements, which are not readily quantifiable or 'solvable' in any objective manner.[116] Such editorial judgements in the mass media have historically been protected against direct government regulation, given the threats to freedom of expression,[117] and attempts to regulate recommendations may raise similar concerns. From this perspective, government attempts to prescribe what is downranked risk becoming a form of censorship—and what is promoted, a form of propaganda.[118] How can a government agency make such essentially political assessments in a legitimate and trustworthy manner?

Put differently, government auditing powers continue to raise issues related to what Kaminski terms 'second-order accountability': is the governance system itself sufficiently open to outside scrutiny?[119] If government determinations rely on privileged access to confidential data, which is not accessible to broader publics, it may be difficult for citizens to scrutinise and contest government policy in this space.

114 Beata Klimkiewicz, 'Is the Clash of Rationalities Leading Nowhere? Media Pluralism in European Regulatory Policies', in Andrea Czepek, Melanie Hellwig and Eva Nowak (eds.), *Press Freedom and Pluralism in Europe: Concepts and Conditions* (University of Chicago Press 2009).

115 Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 66).

116 See Section 2.2 above.

117 e.g. *Jersild v Denmark* [1994] ECtHR 15890/89.

118 In the context of search engines, see: Van Hoboken, *Search Engine Freedom* (n 14).

119 Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 66).

This critique of second-order accountability is in line with constitutional principles on the rule of law, due process and open government, which reflect broad agreement that government action should be documented publicly inasmuch as possible.[120] Also relevant is the Council of Europe's emphasis that oversight of social media recommendations should itself be conducted through 'open' and 'transparent' initiatives.[121] From this perspective, the legitimacy of government action regarding content recommendations depends on its ability to publicise their actions in a meaningful way. However, publicly documenting algorithmic gatekeeping involves significant technical and operational challenges (as discussed in Section 4 below), and has unfortunately not received detailed attention in relevant standards to date.

A final note on *informal* government actions: It is by now well-documented in platform governance that governments can and have used informal means of persuasion and coercion, including the *threat* of regulation, to persuade platforms to adopt certain policies—a stratagem also known as 'jawboning', 'power laundering' or 'regulation by raised eyebrow'.[122] As a result, it can be difficult to disentangle public and private sources of influence in online content moderation; what is presented as a private platform policy may in fact be inspired or compelled by governments, whose role becomes obscured. Indeed, this informal approach is exemplified in the European Commission's ongoing reliance on quasi-voluntary 'Codes'.[123] These informal dimensions of public power risk sidestepping safeguards applicable to formal government action, including transparency principles.[124] In this light, transparency obligations focusing solely on formal government action may fail to capture the full picture. This is where independent disclosure obligations imposed on the platforms may be useful: they may offer a starting point not only for holding the platform itself accountable, but also for detecting and contesting informal government action.

120 On platform governance and rule of law principles, see: Suzor, 'Digital Constitutionalism' (n 70).

121 Council of Europe, CM/Rec(2018)1.

122 Respectively: Derek Bambauer, 'Against Jawboning' (2015) 100 *Minnesota Law Review* 51. Daphne Keller, 'Who Do You Sue? State and Platform Hybrid Power over Speech' (2019) Hoover Institution Aegis Series 1902. <https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf> accessed 15 September 2022. Yochai Benkler, 'A Free, Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate', (2011) 46 *Harvard Civil Rights-Civil Liberties Review* 311. For application to the ECHR context, see: Paddy Leerssen, 'Cut out by the middle man: the free speech implications of social media blocking and banning in the EU' (2015) 6 *JIPITEC* 99.

123 Kuczerawy, 'Fighting online disinformation' (n 91).

124 Tambini, Leonardi and Marsden, 'The privatisation of censorship: self-regulation and freedom of expression' (n 41).

3.3 Research partnerships with academia and civil society

Recent European standards increasingly emphasise the role of independent researchers from academia, civil society, and related categories such as 'the research community' or 'media organisations'. The types of accountability envisaged with these measures are various: in some cases, these actors are formally incorporated in (co-) regulatory decision making processes, and serve clearly designated accountability functions such as fact-checking or regulatory guidance. In other cases, the aims of involving independent researchers appear to be more open-ended, treating independent research and reporting as an end in itself.

A formalised role for civil society actors can be found in in the Council of Europe's 2018 Recommendation on Media Pluralism, which proposes 'open, independent, transparent and participatory initiatives by social media, media actors, civil society, academia and other relevant stakeholders' which would be tasked not only with enabling independent research but also with devising new strategies to ensure diversity and other public interest principles in online content distribution.[125] In France, a 2019 report for the Secretary for Digital Affairs similarly recommends a permanent convening of a 'political dialogue with social networks involving the regulator and civil society', including 'NGOs, regions and the educational and academic communities' with the government tasked with ensuring transparency for the stakeholders involved.[126] Academia and civil society are also increasingly represented in voluntary self-regulatory organs, ranging from the long-standing Global Network Initiative to Facebook's novel and widely-publicised Oversight Board.[127]

More open-ended calls to enable independent research can be found in the EU Code of Practice on Disinformation. Its signatories have committed to 'empower the research community', which includes 'sharing privacy protected datasets, undertaking joint research, or otherwise partnering with academics and civil society organisations if relevant and possible'; and to 'convene an annual event to foster discussions within academia, the fact-checking community and members of the value chain'.[128] In late 2019, the European Commission also issued a call for tenders for a new European Digital Media Observatory, which would allow 'fact-checkers and academic researchers, to

125 Council of Europe, CM/Rec(2018)1.

126 Secretary of State for Digital Affairs of the Republic of France, 'Creating a French framework to make social media platforms more accountable: Final Mission Report on the Regulation of social networks (2019) <https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf> accessed 26 September 2022.

127 Robert Gorwa, 'The platform governance triangle: conceptualising the informal regulation of online content' (2019) 8(2) *Internet Policy Review* 2 <https://doi.org/10.14763/2019.2.1407> accessed 19 September 2022.

128 EU Code of Practice on Disinformation, p. 1.

bring together their efforts and actively collaborate with media organisations and media literacy experts', with the aim to 'fight disinformation online'.[129] To this end, the Observatory would also 'help design a framework to ensure secure access to platforms' data for academic researchers working to better understand disinformation'.[130] The UK's DCMS White Paper would task the government with 'encouraging' the creation of 'access for independent researchers', with the aim of 'ensure that academics have access to company data to undertake research, subject to suitable safeguards' in order to 'help the regulator to assess the changing nature of harms and the risks associated with them.'[131]

Whether it is for independent research or as part of some more formalised co-regulatory process, all of the transparency arrangements in this space tend to emphasise the sharing of data with specific, selected institutions—as 'partners', 'initiatives', or 'observatories'. No explicit attention is paid to creating robust systems of *public* access, available to academia and civil society at large. This preference for partnerships appears to be motivated by the risk of abuse of sensitive data, as highlighted in Cambridge Analytica scandal.[132] By selecting and vetting trustworthy civil society 'partners', and imposing binding conditions and potential sanctions on their access to research data, partnerships have a clear utility in enabling research into sensitive data while reducing the risk of its abuse.

But this selecting and vetting of eligible civil society participants brings challenges of its own. Compared to public datasets, one necessarily reduces the number of stakeholders who can access relevant data and perform research, thereby limiting the potential scale and impact of disclosures. More fundamentally, the selection of eligible participants raises difficult questions about the inclusiveness, diversity and independence of the access framework. The Council of Europe recommends that data access initiatives should be 'open, independent, and participatory', as mentioned previously, but what does this ideal look like in practice? For academia but especially for civil society in a broader sense, there is a very clear tension between these ideals of openness and inclusiveness on the one hand, and the push to restrict access to trusted participants on the other hand. As will be argued below, European governments will face significant challenges in instituting such social media watchdogs—and public access can help to address relevant concerns.

129 European Commission, 'Commission launches call to create the European Digital Media Observatory' (2019) <https://digital-strategy.ec.europa.eu/en/news/commission-launches-call-create-european-digital-media-observatory> accessed 27 September 2022.

130 Ibid.

131 Department for Digital, Culture, Media & Sport, Online Harms White Paper (2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf> accessed 27 September 2022.

132 Freelon, 'Computational research in the post-API age' (n 60).

Academics make promising candidates for research access, not only given their professional expertise and ethical standards, but also due to the university system allowing for a relatively stable and objective means of accreditation (as well as, at the EU level, the European Research Council). Where self-regulatory efforts in this space such as Social Science One been criticised for slow rollout, a lack of (perceived) independence, and a lack of diversity in its leadership, binding regulation could play an important role in addressing relevant concerns and facilitating access for even the most critical research perspectives.[133] This would require neutral and impartial processes for vetting researchers and holding them accountable to data protection and research ethics standards, perhaps in co-regulation with academic institutions themselves. A useful starting point for such efforts is the European Data Protection Supervisor's recent Preliminary Opinion on Data Protection and Scientific Research, which envisages the creation of a co-regulatory accreditation scheme and Code of Conduct for research integrity under the guidance of Data Protection Authorities.[134]

However, despite the clear benefits of such academic research frameworks, they cannot make up for the full breadth of civil society watchdog functions in media governance, which have also been performed by journalists, activists, NGOs and political campaigners. For instance, academia tends to have slow turnover times, and may therefore be ill-equipped to perform real-time, large-scale tasks such as, for instance, fact-checking or election monitoring.[135] Activists, journalists and other civil society actors outside the university system are developing powerful new practices such as algorithmic journalism, platform journalism, and social media activism, which risk being excluded in an academics-only approach to social accountability.[136] Yet for these non-academic institutions, whose membership is often porous and fragmented, defining and accrediting eligible participants is even more fraught. Attempts could be made to devise clear and objective processes for accreditation, which, as in academia,

133 Bruns, 'After the APICalypse' (n 58). Claes de Vreese and others, 'Public statement from the Co-Chairs and European Advisory Committee of Social Science One'. Social Science One (11 December 2019) <https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one> accessed 15 September 2022.

134 European Data Protection Supervisor, 'A Preliminary Opinion on data protection and scientific research' (n 131).

135 Rory Cellan-Jones, 'Facebook's News Feed experiment panics publishers', BBC News (24 October 2017). <https://www.bbc.com/news/technology-41733119> accessed 15 September 2022. More generally, see: Cornia and others, 'Private Sector News, Social Media Distribution, and Algorithm Change' (n 88).

136 On algorithmic journalism as an emerging field, see: Nicholas Diakopoulos, 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures' (2014) 3 *Digital Journalism* 398. Nicholas Diakopoulos, 'The Algorithms Beat: Angles and Methods for Investigation' in Jonathan Gray and Liliana Bounegru (eds.), *The Data Journalism Handbook* 2.0 (Amsterdam University Press 2018).

could interface with existing self-regulatory bodies in the field of journalism such as the European Federation of Journalists. But even this approach may be at once too broad and too narrow: on the one hand, if the goal is indeed to limit disclosures of confidential data to a restricted group of trusted and accountable actors, such broad professional structures might be overly broad and enable abuse. And on the other hand, these professional structures could still be considered too restrictive since they still exclude a range of non-traditional watchdogs such has citizen journalists, activists, influencers, bloggers or NGOs. In essence, what is at stake here is a tension between the practical need to restrict sensitive data access to vetted actors, and conceptions of the fourth estate and civil society as open and participatory institutions.

The difficulties in defining civil society membership are evident in Facebook's self-regulatory attempts to partner with civil society. For instance, Facebook's fact-checking program, which partnered with independent journalists through the Poynter Institutes' International Fact-Checking Network, drew extensive criticism for including the US-based Daily Caller as a partner.[137] This website has been accused by many left-leaning outlets of playing a key role in spreading disinformation and hate speech, arguably invalidating their position as a reliable fact-checker.[138] The point was even raised during CEO Mark Zuckerberg's testimony before Congress.[139] Ultimately, the partnership was terminated in November 2019.[140] A comparable controversy occurred when the Weekly Standard, another right-leaning fact-checking partner, rated an article from the left-leaning ThinkProgress as false.[141] Facebook's widely-publicised Oversight Board is no exception to this trend; the announcement of its membership was met with immediate backlash, mainly consisting of conservative press alleging left-wing bias, but also, for instance, broader concerns about a lack of geographical diversity (e.g. an excess of US members; insufficient Southeast Asian

137 Aaron Rupar, 'Facebook's controversial fact-checking partnership with a Daily Caller-funded website, explained', *Vox* (2 May 2019) <https://www.vox.com/2019/5/2/18522758/facebook-fact-checking-partnership-daily-caller> accessed 10 September 2022.

138 Robert Faris and others, 'Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election' (Berkman Klein Center Research Report 2017) <https://dash.harvard.edu/bitstream/handle/1/33759251/2017-08_electionReport_0.pdf?sequence=9&isAllowed=y> accessed 15 September 2022.

139 Ali Breland, 'AOC Asked Mark Zuckerberg About Facebook's Fact-Checking Process. He Didn't Give Her the Whole Truth' , *Mother Jones* (23 October 2019) https://www.motherjones.com/politics/2019/10/aoc-zuckerberg-facebook-congress-daily-caller-fact-check-dodge/ accessed 15 september 2022.

140 Rupar, 'Facebook's controversial fact-checking partnership with a Daily Caller-funded website, explained' (n 137).

141 Zach Beauchamp, 'Facebook blocked the spread of a liberal article because a conservative told it to', *Vox* (12 September 2018). <https://www.vox.com/policy-and-politics/2018/9/12/17848026/facebook-thinkprogress-weekly-standard> accessed 16 September 2022.

members).[142] Such cases illustrate the difficulty, amidst the ongoing decline in trust in mainstream media and established knowledge institutions, of arriving at generally accepted definitions and configurations of 'civil society'.

It is worth noting that the regulation of civil society research access is in a less advanced stage compared to user-facing disclaimers and government oversight; it is largely limited to soft law, and no binding legislation has yet been proposed in this space. It remains to be seen whether and how relevant legislators will take up their cause in upcoming rounds of legislation. Existing standards do indicate, however, a focus on vetted partnerships for privileged data access, as opposed to the furnishing of publicly accessible information.

In sum, the push to create exclusive research access programs has important advantages in enabling in-depth investigative work, but also has limitations. There are important trade-offs between the vetting of eligible researchers for sensitive data access, on the one hand, and the potential scale, diversity and independence of such programs on the other hand. Forthcoming plans for regulated research access will require careful attention to institutional design so as to manage such trade-offs, and ensure the independence and credibility of these research efforts. Overall, these confidential access programs will have much to offer for in-depth academic research, but appear less suitable for real-time monitoring and reporting by journalists, activists and other non-academic civil society actors.

4. The case for public access

The above has shown that the emergent European framework for transparency in social media recommendations focuses on channelling information towards user-facing notices, government authorities, and civil society research partners. In this landscape, preciously little information is made publicly available. Independent observation of personalised recommendations is obstructed by their technical and legal design. User-facing disclosures, whilst public, are typically simplified and individualised. Detailed data is accessible only to a privileged few in government and selective research partnerships.

142 e.g. Robin Pagnamenta, 'Facebook will rue its left-wing oversight board appointments', *The Telegraph* (6 May 2020) <https://www.telegraph.co.uk/technology/2020/05/06/facebook-will-rue-left-wing-oversight-board-appointments/> accessed 19 September 2022. Jenny Domino, 'Why Facebook's Oversight Board is Not Diverse Enough', *Just Security* (21 May 2020) <https://www.justsecurity.org/70301/why-facebooks-oversight-board-is-not-diverse-enough/ accessed> accessed 15 September 2022.

A robust regime for public access, I argue, would contribute not only to the first-order goal of making oversight of platforms more effective, but also to the second-order goal of making the governance system as a whole more open to outside critique.[143] This section articulates these potential benefits associated with public access, and suggests some starting points for its design and regulation. In particular, these recommendations focus on the automated, real-time disclosure of high-level, anonymised information about recommendation system outputs, audiences, and organisations.

4.1 The pros and cons of public access

The main drawback of public records, compared to confidential disclosures such as data sharing partnerships and government auditing, is their limitations in sharing sensitive data: public access requires a trustless design that pre-empts abuse by malicious actors. In the context of platform recommender systems, disclosures would need to contend with threats to user privacy, and, according to platforms, the integrity of the service (i.e. by enabling third parties to 'game' the algorithm).[144] Privacy-by-design techniques such as anonymisation and differential privacy can go some way in mitigating these concerns. Nonetheless, publicity places hard limits on what can be disclosed and thus on the ultimate research utility of public disclosures.

However, public access also has important advantages over data partnerships in terms of increasing inclusiveness and scalability. By simply making information publicly accessible, one side-steps the pitfalls of needing to define, accredit or otherwise institutionalise such factious and amorphous categories as 'civil society' or 'academia'. Public disclosures would be available to every researcher with the time and interest—not the lucky few with the wherewithal and *bona fides* to engage in protracted negotiations, tender procedures or other forms of partnership arrangements. In particular, public access opens the doors to civil society actors that do not have an institutional means of accreditation, such as many independent journalists, NGOs, activists, and so forth. In this way, public records offer the prospect of broader and more diverse uptake. Public access can also mitigate threats to researchers' independence, both real and perceived, since it leaves its users free to pursue critical lines of research without needing to appease the purveyors of data—be they platforms in as self-regulatory setting, or governments in a regulated setting. In this way, public records avoid many of the aforementioned problems with more institutionalised 'partnership' models for data access.

143 This distinction between first- and second-order accountability is drawn from Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 66).

144 Burrell, 'How the machine thinks' (n 47). Pasquale, *The Black Box Society* (n 3).

The above suggests that public records could be instrumental for real-time, high-level monitoring by media watchdogs such as academics, journalists, activists, and NGOs. As discussed in Section 3.3, academics and others performing more in-depth research may be relatively well-served by privileged research partnerships. But even here it is worth considering that public records can offer a low-cost starting point for more in-depth research. For instance, public records may not suffice to conclusively demonstrate bias or discrimination in a recommendation algorithm, but at a minimum they can offer a starting point for such inquiries by rendering visible trends and disparate outcomes in the system's recommendation outputs. Such evidence can then be used to request further clarification from the platform,[145] or investigate with more fine-grained tools, such as algorithmic auditing approaches,[146] data surveys[147] or GPDR data access requests.[148] In other words, public access can serve as a first-warning system for more targeted efforts.

The open nature of public disclosures means that they can also contribute to the second-order goal of holding the governance system *itself* accountable—i.e. Kaminski's ideal of 'second order accountability'. As discussed, direct government regulation of social media recommendations is problematic from a fundamental rights perspective, since it applies opaque, technocratic methods to a highly contentious and politically sensitive field of governance. Even if multistakeholder perspectives from civil society or academia are incorporated in relevant oversight structures, such institutions run the risk of capture or bias. Releasing public information about content recommendation trends can help to critique such governance structures, and potentially even provide a starting point to identify more informal 'jawboning' relationships between platform power and public power.

A similar argument about second-order accountability may also apply to other platform *users*, insofar as they also co-determine harmful outcomes in recommender systems. For instance, the discovery that certain harmful channels are being disproportionately recommended towards children could prompt intervention not only from platforms or from governments, but might also appeal to the responsibilities of the content

145 James Grimmelman and Daniel Westreich, 'Incomprehensible Discrimination', 7 *California Law Review Online* <https://scholarship.law.cornell.edu/facpub/1536/> accessed 12 September 2022.

146 Christian Sandvig and others, 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms' (2014), Paper presented to *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* <https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf> accessed 26 September 2022.

147 Eduardo Hargreaves and others, 'Fairness in Online Social Network Timelines: Measurements, Models and Mechanism Design', (2019) 127 *Performance Evaluation Review* 15.

148 Jef Ausloos, 'GDPR Transparency as a Research Method' (2019) SSRN Working paper. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3465680> accessed 16 September 2022.

provider in question. Public disclosures could ideally accelerate such media criticism by providing the necessary evidence of relevant recommendation trends. In sum, then, public records can assist civil society actors in holding not only platforms accountable but the governance system as a whole.

The benefits of public access pertain not only to civil society, but may also spill over to government oversight. Since regulatory auditing and other investigative powers can be slow and costly to perform, publicly available data can cut down on such costs and help agencies to perform high-level monitoring and more efficiently prioritise their in-depth investigative efforts. Perhaps more important, however, is the earlier point that regulatory enforcement commonly relies on knowledge sourced from third parties, e.g. through consultation responses, complaints, tips, and referrals, and scientific literature.[149] In this light, public data access can also redound to the benefit of government regulation, by helping third parties in monitoring social media recommendation systems, and referring cases to competent government agencies. For instance, content providers who depend on platforms to disseminate their content have incentives to monitor recommendations trends and check for potentially unlawful or anticompetitive patterns of discrimination. Public records could help them in such efforts, whereas government-focused transparency places the onus entirely on the government to do its own monitoring.

Of course, all of these potential benefits related to public access are still largely speculative, and their realisation depends on whether it is implemented effectively so as to offer meaningful and accessible information. The operational and technical challenges in designing such a regime are not to be underestimated, and further research is needed to pre-empt possible abuses. The above is simply intended to articulate the distinct benefits of public access, relative to more exclusive approaches currently seen in Europe policy. These benefits are particularly salient, it is submitted, in the politically sensitive context of media governance, where scepticism of both market and public ordering are uniquely strong and the demands of broad and inclusive second-order accountability are therefore particularly urgent.

4.2 Designing public access

A long line of transparency research has emphasised that transparency measures must be designed with the needs of their intended users in mind.[150] So what information, specifically, requires public access? This is a complex question, particularly since the potential userbase for public disclosures is necessarily undefined and open-ended,

149 Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 66).

150 Fung, Weil and Graham, *Full Disclosure* (n 76).

leaving it outside the scope of this paper to offer an exhaustive answer. Focusing on the needs of academia and civil society in particular, what follows is intended as exploratory, offering some starting points for further research and debate. In terms of format, public disclosures about recommender systems should include real-time, high-level, anonymized data access through public APIs and browser interfaces. In terms of content, public access should cover the documentation of recommendation outputs and their audiences; content-specific ranking decisions and other interventions by the operator's in recommendation system performance; and the organisational structures that control recommendation systems.

At the outset, an important starting point in terms of existing best practices is public research APIs. As discussed in Section 2.3, many social media platforms already offer some level of real-time public access through these systems, and public regulation can draw and build on this prior art. The functionality of these systems has been reduced significantly in recent years, nominally in response to privacy concerns resulting from the Cambridge Analytica scandal, but communications researchers argue that there is evidence of a disproportionate overreaction, and the pendulum has now swung too far back from openness to secrecy.[151] Binding public regulation could provide an impetus for (privacy-compliant) reform. To this end, policymakers can draw on expertise from communications science and adjacent fields, which command extensive experience with the design and usage of such public APIs.

As for the substance of public disclosures about content recommendations, one particularly salient aspect of their design which could be eligible for disclosure is content-specific ranking interventions. Platforms routinely intervene in recommender systems to alter specific outcomes, and such information could be eligible for disclosure. For instance, as mentioned in Section 2.2, Facebook currently partners with third-party fact-checkers to identify and downrank 'false headlines' from untrustworthy news sources, and these fact-checkers publish explanations for each intervention they make. A more ambitious approach would register such decisions in a central platform repository, rather than dispersing them across various partnered websites. Ideally, such an approach would not only apply to third-party fact-checkers but to all human interventions in the algorithmic ranking system across the board, whether by platform workers or external partners. Such public records need not require full disclosure of the recommendation algorithm as a whole, as this could undermine service integrity and enable gaming of these systems by spammers and other malicious actors.[152]

151 See Section 2.3 above.

152 e.g. Pasquale, *The Black Box Society* (n 3).

An instructive comparison can be made between downranking and content removal decisions. For content removal decisions, platforms have declined content-level disclosure because the content at issue is by its very nature expected to be illegal or otherwise unsuitable for publication.[153] But this rationale does not apply when it comes to downranking decisions, since these are expressly intended for content that platforms do *not* wish to remove. In this light, there appears to be no compelling reason why these downranking decisions should not be made a matter of public record.

More broadly, it is worth investigating a best effort documentation requirement for other human-coded aspects of recommender algorithms. While many aspects of these algorithms are the product of complex machine-learning processes and therefore difficult to understand or explain even for their makers, other elements are human-coded and therefore easier to shed light on. One example is Facebook's Click-Gap initiative, which identifies low-quality based on the ratio of engagement on Facebook versus overall popularity across the web and thus serves to privilege more 'mainstream', established media outlets.[154] It is to Facebook's credit that this update has been announced publicly.[155] But this is arguably the exception proving the more fundamental rule that conscious and explainable interventions are taking place, without any guarantee that these are necessarily disclosed to the public. How else have platforms intervened to curate their recommendations? Indeed, as discussed, YouTube boasts that it has made 'hundreds' of changes in 2018 alone, and it is unclear what these entail.[156] A legal requirement that such interventions in the algorithm must be disclosed systematically would help to prevent any important omissions and underwrite the significance of platform disclosures. Of course, an important limitation is the risk of gaming in the algorithm, which may counsel against overly detailed specification of such changes: for instance if the specific keywords of a an anti-spam blacklist were to be disclosed. Such defences may need to be evaluated on a case-by-case basis.

Public documentation of recommender systems need not focus exclusively on the *algorithm*. As discussed in Section 2.3, the algorithm as such cannot account fully for the effects and outcomes of recommender systems, as this requires reference to users' and their activity. As argued by Rieder, Matamoroz-Fernandez, and Coromina

153 q.v. Heidi Tworek and Paddy Leerssen, 'An Analysis of Germany's NetzDG Law' (Transatlantic High Level Working Group on Content Moderation Research Report 2019) <https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf> accessed 19 September 2020.

154 See Section 2.2 above.

155 Rosen, 'Remove, Reduce, Inform: New Steps to Manage Problematic Content' (n 38).

156 YouTube, 'Continuing our work to improve recommendations on YouTube' (n 39).

argue, this can best be approached through the study of recommendation *outcomes*, in terms of what content is recommended, and to whom.[157] At present, social media recommendations are scarcely documented for many important platforms. What are the most recommended pieces on content on YouTube or Instagram on a given day? In a given country? For a given age group? Some knowledge can be gleaned through independent observational methods, but, as discussed in Section 2.3, such methods face major operational challenges and necessarily produce incomplete and time-lagged datasets. Perhaps the most ambitious project in this space, Algotransparency.org, only covers YouTube recommendations made by 1000 selected channels and on a limited set of keywords.[158] While such methods have already led to important insights about social media recommendations,[159] far more comprehensive and systematic data could be published with the (regulated) cooperation of platforms themselves.[160] In essence, these output disclosures would serve to recreate, to some degree, the baseline publicity or visibility that was inherent in mass media content distribution, and has been lost through personalisation. Even if the precise motives and decisions of our gatekeepers remains secret, at least the overall outcomes in terms of content distribution can then start to be observed.

Such public documentation of outputs would require strong safeguards against potential privacy harms. One important best practice is to limit disclosures to *publicly accessible* content—as opposed to private content such as personal messages. But this is no panacea: even though such an API would not technically expand access to content, since the content is already public, it would still make the content searchable and measurable in new ways for third parties, which may raise privacy issues of its own.[161] In addition, therefore, public access can also apply a *de minimis* rule: only content above a certain threshold of popularity could be included. Such a rule, already commonly seen in public research APIs, would limit the scope to the most important and visible content, and protect more sensitive activity from excessive monitoring. Potentially,

157 Rieder, Matamoroz-Fernandez and Coromina, "From ranking algorithms to "ranking cultures"" (n 10).

158 n.a., *AlgoTransparency* (n.d.) <https://www.algotransparency.org/> accessed 27 September 2022.

159 e.g. Paul Lewis, 'Fiction is outperforming reality': how YouTube's algorithm distorts truth', *The Guardian* (2 February 2018) <https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth> accessed 19 September 2022. Guillaume Chaslot, 'YouTube's A.I. was divisive in the US presidential election', *Medium* (27 November 2016). <https://medium.com/the-graph/youtubes-ai-is-neutral-towards-clicks-but-is-biased-towards-people-and-ideas-3a2f643dea9a> accessed 15 September 2022.

160 Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' (2019) 10(3) *European Journal of Law and Technology* <https://ejlt.org/index.php/ejlt/article/view/686> accessed 15 September 2022.

161 Michael Zimmer, '"But the data is already public": on the ethics of research on Facebook' (2010) *12 Ethics in Information Technology* 313.

content below this threshold could still be described with certain metadata, such as keywords, format, or language, to provide at least some basic insight into content flows without threatening the underlying content.

In designing public access regimes and their privacy safeguards, a relevant precedent is the experience with political advertising archives. Since 2018, major social media platforms have started developing such public archives to document political ads sold on their services.[162] Like the output documentations discussed in this paper, ad archives are similarly concerned with accountability in the algorithmically personalised distribution of content—albeit in the particular context of advertising content. In ad archives, the data is disclosed through searchable web interfaces as well as through APIs, and audience data are anonymised and aggregated to avoid user privacy concerns. It should be noted that present self-regulatory implementations have been criticised extensively; for instance, researchers from Mozilla concluding that the API was so bug-ridden as to be effectively unusable—a strong argument for regulation in this space.[163] Nevertheless, whilst their research utility is necessarily limited for understanding deeper questions about algorithmic sorting and bias, platform ad archives have already started to see regular use in real-time media monitoring and election coverage.[164] In this light, the experience with ad archives is instructive in two ways. First, it warns against an overreliance on self-regulation, given the critical failures of voluntary initiatives in this space. Second, despite their inadequate implementation in practice, ad archives do provide a basic conceptual blueprint for public transparency in algorithmic content distribution: real-time, anonymised, output-focused, and accessible to all.

A final point of attention for public access is the organisations behind recommender systems. Information about these organisations is highly relevant for understanding how gatekeeping decisions are made, and better outcomes can be ensured. Relevant issues in the context of recommender systems could include the location, demographic background, training, reward schemes, authorisations, and management systems in place for relevant workers. Comparable rules about organisational transparency can

162 Chapter 3 below (Paddy Leerssen and others, 'Platform ad archives: promises and pitfalls' (2019) 8(4) Internet Policy Review <https://policyreview.info/articles/analysis/platform-ad-archives-promises-and-pitfalls> accessed 1 September 2022).

163 Matthew Rosenberg, 'Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised', *The New York Times* (25 July 2019) <https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html> accessed 19 September 2022.

164 See, for instance: Julia Wong, 'One year inside Trump's monumental Facebook campaign', *The Guardian* (28 January 2020) <https://www.theguardian.com/us-news/2020/jan/28/donald-trump-facebook-ad-campaign-2020-election?CMP=Share_iOSApp_Other> accessed 20 September 2020.

already be found in Germany's *Netzwerkdurchsetzungsgesetz*, which includes public documentation requirements for staffing and training of content removal operations related to this law.[165] Professional standards for transparency in journalistic organisations can also serve as a template.

4.3 Regulating public access

Like most forms of platform transparency, public access will require a binding legal basis in order to be effective. As discussed in Section 2.3, platforms have a poor track record in their voluntary transparency reforms, and even when cooperating in earnest may still be dogged by questions of credibility and independence. Binding transparency obligations can help to address these concerns, and avoid a situation in which 'only approved questions get answered'.[166] In addition, public regulation of transparency can help to offset legal restrictions on data disclosure, e.g. by providing relevant exemptions under intellectual property law, or processing grounds under data protection law.

Another advantage of binding regulation is that it may remedy the current precarity of automated research access, discussed in Section 2.3. Bernhard Rieder and Jeannette Hofmann argue that the goal in platform governance should be to 'transpos[e] local experiments into more robust practices able to guarantee continuity and accumulation', leading to 'structured interfaces between platforms and society'.[167] Relevant to this endeavour is Fung et al's research on the 'sustainability' of transparency measures, which recommends that they improve in scope, accuracy, and use over time.[168] With this in mind, it is advisable for a regulatory effort to start with a relatively modest scale, perhaps as a pilot study or experiment, and then gradually expand in respond to feedback from early users.[169] Other elements of sustainable transparency highlighted by Fung et al include effective enforcement of applicable rules, and the strengthening of potential user groups such as civil society organisations.[170]

165 Tworek and Leerssen, 'An Analysis of Germany's NetzDG Law' (n 153). Wagner and others, 'Regulating Transparency?' (n 70).

166 Brother Ali (2007), 'Uncle Sam Goddamn'. *The Undisputed Truth* [CD]. Minneapolis: Rhymesayers Entertainment.

167 Bernhard Rieder and Jeanette Hofmann, 'Towards Platform Observability'(2020) 9(4) *Internet Policy Review* <https://doi.org/10.14763/2020.4.1535> accessed 19 September 2022

168 Fung and others, *Full Disclosure* (n 76).

169 Jef Ausloos, Paddy Leerssen and Pim ten Thije, 'Operationalizing Research Access in Platform Governance: What To Learn From Other Industries?' (Research Report AlgorithmWatch 2020). <https://algorithmwatch.org/en/governing-platforms-ivir-study-june-2020/#study> accessed 16 September 2020.

170 Fung and others, *Full Disclosure* (n76, describing civil society organisations as 'information intermediaries', who mediate between the disclosing party and the ultimate end users of information)

Due to platform dominance dynamics,[171] size-based regulation is appropriate; targeting the most influential platforms addresses the major sources of risk while avoiding unnecessary or disproportionate burdens on smaller services. For instance, platform size could be defined based on revenue, user count, view count, or some combination of these metrics. Similar size-based regulation is already common in recent proposals for transparency in social media platforms, such as the EU Code of Practice, the US Honest Ads Act and Germany's *Netzwerkdurchsetzungsgesetz*.

Given the complexity of designing privacy-compliant disclosure standards, rules for public access will be difficult to codify exhaustively in one-size-fits-all legislation. Not only are social media recommendations technically complex, they are also heterogeneous; each has unique features (types of posts and formats, engagement metrics, et cetera) which may require unique forms of documentation and privacy safeguards. In response, disclosure standards in legislation must remain broad, to be specified in case-by-case by an authorised regulator. Of course, such a body could also be instrumental in achieving other transparency goals in social media governance besides public access, such as those related to individual user notices, regulatory enforcement and exclusive research access frameworks.

An ongoing challenge regarding transparency regulation is finding an appropriate regulatory body to enforce these rules. Few national systems have developed agencies equipped to regulate social media, and leaving Member States to each develop their own institutional capacities risks not only a duplication of efforts but also the risk of regulatory fragmentation and potentially conflicting standards. Transparency measures, like most information products, tend to benefit from economies of scale, which support the case for uniform regulation at EU level. And yet, a (social) media regulator does not yet exist at the EU level. Indeed, member states have historically resisted the creation of a EU media regulator given the cultural and political sensitivities in this space.[172]

Whilst developing a definitive division of competences is outside the scope of this paper, it is worth emphasising that the regulation of transparency may in theory be separated from substantive media policy. Under such an approach, the EU could put its full force behind ensuring access to information, whilst leaving national entities to make use of this data for their various regulatory efforts and thus to realise the *substance* of domestic media policy.[173] This aligns with other proposals to institute government regulators focused on transparency. For instance, Ben Wagner and

171 Barwise and Watkins, 'The evolution of digital dominance: how and why we got to GAFA' (n 171).
172 Klimkiewicz, 'Media Pluralism in European Regulatory Policies' (n 114).
173 Ibid.

Lubos Kuklis envisage a 'single European institution which could act as an auditing intermediary to ensure that the data provided to regulators by social media companies are accurate'.[174] Their proposal focuses on transparency towards other public regulators, such as data protection and competition authorities, but a similar vision could also apply to public access and its use by a range of governance stakeholders in government and civil society. The ideas about public access regulation outlined in this paper should be considered alongside such broader debates about the need for dedicated regulatory structures for transparency for platforms in general, and social media in particular.

5. Conclusion

These are decisive times for the regulation of social media content recommendations. As the 'techlash' moves from opinion pages to public policy, and attempts at regulation begin in earnest, we see a variety of attempts to make social media platforms more transparent and accountable in their content recommendations. A governance landscape is emerging in which users, governments and civil society all have a role to play in holding these systems accountable, and realising public values in our content feeds. Transparency rules are developing accordingly, with each stakeholder group being associated with its own types of disclosures. As recurring themes in ongoing policy, this paper has identified notices and disclaimers, government auditing, and data access partnerships.

A central component in ongoing efforts is the enabling of independent oversight by academia and civil society. This is laudable given the particular sensitivity of recommender governance from the perspective of democracy and fundamental rights. Yet this paper has cautioned against efforts which pursue transparency towards academia and civil society exclusively through institutionalised systems of privileged data access. Whilst such privileged access regimes have important advantages in enabling in-depth scholarly research, there may be low-hanging fruit of non-sensitive data that could find far wider uptake if made public without restriction. This paper has articulated how real-time, high-level public access has distinct advantages for accountability in this space. A robust system of public access not only allows for wider uptake and greater impact, but is essential to make the technocratic, expert-driven institutions of recommendation governance accountable to scrutiny and contestation by broader publics and interest groups.

174 Ben Wagner and Lubos Kuklis (2019), 'Disinformation, data verification and social media', *Media@LSE* (7 January 2020) <https://blogs.lse.ac.uk/medialse/2020/01/07/disinformation-data-verification-and-social-media/> accessed 20 September 2020.

This paper has also provided some starting points for the design and regulation of such public access. Overall, it suggests a reorientation from 'algorithms' as objects of transparency towards a broader inquiry into the sociotechnical dynamics of recommender systems. To this end, fruitful avenues for public access include content-level detail on downranking decisions and other manual interventions in the recommender system, as well as publicly searchable documentation of recommendation outputs for the most popular content. More generally, policymakers should explore existing best practices in the design and regulation of public research APIs. Given the complexity of these issues, the most promising way to regulate this would be through broad legislative standards, specified and enforced by an authorised regulator. This approach resonates with recent proposals in academia and government to install a dedicated transparency regulator for online platforms.

On a final note: This paper's discussion of transparency has hewed closely to the prevailing vision in European media governance of social media platforms as regulated oligopolists, whose dominance as online speech infrastructure is not to be replaced or contested but rather to be made more transparent and accountable to public interest considerations. It remains to be seen, given the vast power and complexity of these services and the sensitivity of the data they process, whether such a vision can be realised. We may well come to conclude that for-profit stewardship of these influential and opaque systems simply creates unacceptable and unmanageable risks to democracy; that meaningful transparency in these circumstances is a false hope—much less accountability—and that instead more fundamental changes to ownership or business models may be necessary, such as switching to cooperatively owned or publicly owned social media services.[175] But even for these more radical visions of online media governance, the arguments discussed in this paper may still hold some relevance: what will likely remain are the importance of broad and inclusive scrutiny of algorithmic gatekeeping, and the distinct benefits of publicly accessible information to that end.

175 Evgeny Morozov (2019), 'Digital Socialism? The Calculation Debate in the Age of Big Data'. *New Left Review* 116. Trebor Scholz, *Platform Cooperativism: Challenging the Corporate Sharing Economy* (Rosa Luxemburg Stiftung 2016). Ethan Zuckerman, 'The Case for Digital Public Infrastructure', *Knight First Amendment Institute* (17 January 2020) <https://knightcolumbia.org/content/the-case-for-digital-public-infrastructure> accessed 20 September 2020. Mariana Mazzucato, 'Let's Make Data Into A Public Good', *MIT Policy Review* (27 June 2018) <https://www.technologyreview.com/s/611489/lets-make-private-data-into-a-public-good/> accessed 20 September 2020.

CHAPTER 3

Platform ad archives: promises and pitfalls[1]

1 Originally published as: Paddy Leerssen, Jef Ausloos, Brahim Zarouali, Natali Helberger and Claes de Vreese, 'Platform ad archives: promises and pitfalls' (2019) 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1421>.

Abstract

This paper discusses the new phenomenon of platform ad archives. Over the past year, leading social media platforms have installed publicly accessible databases documenting their political advertisements, and several countries have moved to regulate them. If designed and implemented properly, ad archives can correct for structural informational asymmetries in the online advertising industry, and thereby improve accountability through litigation and through publicity. However, present implementations leave much to be desired. We discuss key criticisms, suggest several improvements and identify areas for future research and debate.

1. Introduction

In 2018, the online platforms Google, Facebook and Twitter all created political ad archives: publicly accessible databases with an overview of political advertisements featured on their services. These measures came in response to mounting concerns over a lack of transparency and accountability in online political advertising, related to illicit spending and voter manipulation. Ad archives have received widespread support in government and civil society. However, their present implementations have also been criticised extensively, by researchers who find their contents to be incomplete or unreliable.[2] Increasingly, governments and civil society actors are therefore setting up their own guidelines for ad archive architecture—in some cases even binding legislation. Ad archive architecture has thus rapidly gained relevance for advertising law and policy scholars, both as a tool *for* regulation and as an object *of* regulation.[3]

2 Mozilla, 'Facebook and Google: This is What an Effective Ad Archive API Looks Like', *The Mozilla Blog* (27 March 2019) <https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like> accessed 19 September 2022. J. Nathan Matias, Austin Hounsel and Melissa Hopkins, 'We Tested Facebook's Ad Screeners and Some Were Too Strict' *The Atlantic* (2 November 2018) <https://www.theatlantic.com/technology/archive/2018/11/do-big-social-media-platforms-have-effective-ad-policies/574609/> accessed 19 September 2022. Jeremy Merrill, 'How Big Oil Dodges Facebook's New Ad Transparency Rules', *ProPublica* (1 November 2018) <https://www.propublica.org/article/how-big-oil-dodges-facebooks-new-ad-transparency-rules> accessed 19 September 2022. Aaron Rieke and Miranda Bogen, Leveling the Platform: Real Transparency for Paid Messages on Facebook. (Research Report UpTurn 2018) <https://www.upturn.org/static/reports/2018/facebook-ads/files/Upturn-Facebook-Ads-2018-05-08.pdf> accessed 19 September 2022. Laura Edelson and others, 'An Analysis of United States Online Political Advertising Transparency' (2019) *ArXiv [Cs]* http://arxiv.org/abs/1902.04385 accessed 15 September 2022. Peter Andringa, 'Interactive: See Political Ads Targeted to You on Facebook' *NBC (16 January* 2018). <http://www.nbcsandiego.com/news/tech/New-Data-Reveal-Wide-Range-Political-Actors-Facebook-469600273.html> accessed 16 September 2022. Izzy Lapowsky, 'Obscure Concealed-Carry Group Spent Millions on Facebook Political Ads', *WIRED* (19 January 2018) <https://www.wired.com/story/facebook-ads-political-concealed-online/> accessed 19 September 2022. O'Sullivan D, 'What an anti-Ted Cruz meme page says about Facebook's political ad policy', CNN (25 October 2018) <https://www.cnn.com/2018/10/25/tech/facebook-ted-cruz-memes/index.html> accessed 19 September 2022. Jim Waterson, 'Obscure pro-Brexit group spends tens of thousands on Facebook ads', *The Guardian* (14 January 2019) <https://www.theguardian.com/politics/2019/jan/14/obscure-pro-brexit-group-britains-future-spends-tens-of-thousands-on-facebook-ads> accessed 20 September 2020. Jonathan Albright, 'Facebook and the 2018 Midterms: A Look at the Data – The Micro-Propaganda Machine' (4 November 2018). <https://medium.com/s/the-micro-propaganda-machine/the-2018-facebook-midterms-part-i-recursive-ad-ccountability-ac090d276097> accessed 15 September 2022. Philip Howard and others, 'The IRA, Social Media and Political Polarization in the United States, 2012–2018' (Research Report Oxford Computational Propaganda Project 2018) < https://www.oii.ox.ac.uk/news-events/reports/the-ira-social-media-and-political-polarization-in-the-united-states-2012-2018/> accessed 15 September 2022.

3 Derived from the general distinction between the governance of platforms and the governance by platforms: Tarleton Gillespie, 'Governance of and by platforms', in: Jean Burgess, Alice Marwick and Thomas Poell (eds), *The SAGE handbook of social media* (Sage 2017).

This article offers an overview of the ad archive governance debate, discussing the potential benefits of these tools as well as pitfalls in their present implementations. Section 2 starts with a basic conceptual and legal framework which describes the basic features of archives and applicable regulations, followed by a normative framework which discusses the potential benefits of ad archives in terms of transparency and accountability. Section 3 reviews the shortcomings of current ad archive initiatives, focusing on three core areas of ongoing debate and criticism. Firstly, we discuss scoping: ad archives have faced difficulty in defining and identifying, at scale, what constitutes a 'political advertisement'. Secondly, verifying: ad archives have proven vulnerable to inauthentic behaviour, particularly from ad buyers seeking to hide their true identity or the origin of their funding. Thirdly, targeting data: ad archives do not document in meaningful detail how ads are targeted or distributed. We propose several improvements to address these shortcomings, where necessary through public regulation. Overall, we argue that both legal scholars and communications scientists should pay close attention to the regulation of, and through, this novel and potentially powerful tool.

2. Promises: the case for ad archives

2.1 Conceptual framework: what are 'ad archives'?

This paper focuses on ad archives, which are systems for the automated public disclosure of advertisements via the internet. The key examples are Facebook's Ad Library, Google's Advertising Transparency Report and Twitter's Ad Transparency Center. These systems document the advertisement messages sold on the platform, as well as associated metadata (e.g., the name of the buyer, the number of views, expenditure, and audience demographics). These archives are public, in the sense that they are available without restriction to anyone with a working internet connection.

In practice, the major ad archives have focused on documenting political advertisements, rather than commercial advertisements. Beyond this, they differ in important respects. Firstly, they differ significantly in how they define 'political' advertising in order to determine what ads are included in the archive. The major archives also differ in how they verify their contents—particularly the identity of their ad buyers—and in terms of the metadata they publish related to ad targeting. Section three considers these questions of scoping, verifying and targeting in further detail.

The major ad archives went live in 2018. Facebook's archive was first announced in October 2017 and went live the next year in May 2018. Google and Twitter followed soon after. They initially focused exclusively on the United States, but they have since

gradually expanded their efforts. Facebook and Twitter's archives now offer worldwide coverage, although certain functions are still regionally restricted. Google covers only the US, the European Union and India.[4]

In theory, ad archives can be created not only by platform intermediaries but also by a range of other actors, including advertisers, academics or NGOs. For instance, political parties can maintain their own online database documenting their political advertisements, as has been proposed in the Netherlands.[5] As early as 2012, Solon Barocas argued for a centralised non-profit database, or 'clearing house', for political ads.[6] The London School of Economic's Truth and Trust Commission proposes that the government administer a central database, or 'political advertising directory'.[7] The investigative journalists of ProPublica have maintained a public database of Facebook ads which they crowd-sourced from a group of volunteers.[8] While we do not discount these approaches, our discussion focuses on platform-operated archives, since these have recently attracted the most widespread traction in policy and practice.'

2.2 Legal framework: why are platforms building archives?

Formally speaking, the major platform ad archives are self-regulatory measures. But they emerged in response to significant public pressure from the ongoing 'techlash'.[9] These 'voluntary' efforts are therefore best understood as an attempt to stave off binding regulation.[10] Indeed, platforms have no immediate commercial incentive to offer transparency in their advertising practices. The role of public regulation, or at

4 Google, 'Verification for election advertising in the European Union' (n.d.) <https://support.google.com/adspolicy/answer/9211218> accessed 26 September 2022.

5 Netherlands Ministry of the Interior, 'Response to the Motion for Complete Transparency in the Buyers of Political Advertisements on Facebook' (2019) Kamerstuk 35 078 Nr 26 <https://www.tweedekamer.nl/kamerstukken/detail?id=2019Z03283&did=2019D07045> accessed 27 September 2022.

6 Solon Barocas, 'The Price of Precision: Voter Microtargeting and Its Potential Harms to the Democratic Process' (2012) *Proceedings of the First Edition Workshop on Politics, Elections and Data* 31.

7 Sonia Livingstone and others, Tackling the Information Crisis: A Policy Framework for Media System Resilience (Report of the LSE Commission on Truth, Trust and Technology 2018) <http://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf> accessed 19 September 2020.

8 Jeremy Merrill and Ariana Tobin, 'Facebook Moves to Block Ad Transparency Tools —Including Ours', *ProPublica* (28 January 2019) <https://www.propublica.org/article/facebook-blocks-ad-transparency-tools> accessed 19 September 2022.

9 Ben Zimmer, 'Techlash': Whipping Up Criticism of the Top Tech Companies', *The Wall Street Journal* (10 January 2019) <https://www.wsj.com/articles/techlash-whipping-up-criticism-of-the-top-tech-companies-11547146279> accessed 24 September 2022.

10 Ben Wagner, 'Free Expression? Dominant information intermediaries as arbiters of internet speech'. In: Martin Moore and Damian Tambini (eds), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (Oxford University Press 2018).

least the threat thereof, is therefore essential in understanding the development of ad archives.[11] Below we offer an overview of key policy developments.

Both platforms and policymakers present ad archives as a means to improve accountability in online political advertising.[12] Political advertising in legacy media has historically been regulated in various ways, to prevent undue influence from concentrated wealth on public discourse. Online advertising is placing new pressure on these legacy regimes. In many cases, the language of existing law has simply not been updated to apply online. Furthermore, online political micro-targeting has unique affordances that can enable new types of harms demanding entirely new regulatory responses. For instance, platform advertising services lower the barrier to buying ads across borders, and to buy ads under false or misleading identities. Furthermore, micro-targeting technology, which enables advertisers to target highly specific user segments based on personal data analytics, can enable novel methods of voter deception, manipulation and discrimination.[13] For instance, targeted advertising can enable politicians to announce different or even conflicting political programmes to different groups, thereby fragmenting public discourse and making it more difficult to hold politicians accountable to their electoral promises.[14] Targeted advertising can also enable discrimination between voter groups, both intentionally through advertisers' targeting decisions and unintentionally through undocumented algorithmic biases.[15]

11 Eveline Vlassenroot and others, 'Web archives as a data resource for digital scholars' (2019) 1 *International Journal of Digital Humanities* 85.

12 e.g. Rob Goldman, 'Update on Our Advertising Transparency and Authenticity Efforts', *Facebook Newsroom* (27 October 2017). <https://newsroom.fb.com/news/2017/10/update-on-our-advertising-transparency-and-authenticity-efforts/> accessed 19 September 2022. Mark Warner, The Honest Ads Act: a primer (US Senate 2017) <https://www.warner.senate.gov/public/index.cfm/the-honest-ads-act > accessed 27 September 2022.

13 Frederik Zuiderveen Borgesius and others, 'Online Political Microtargeting: Promises and Threats for Democracy' (2018) 14 *Utrecht Law Review* 82. Jeff Chester and Kathy C Montgomery, 'The role of digital marketing in political campaigns' (2017) 6(4) *Internet Policy Review* <https://policyreview.info/articles/analysis/role-digital-marketing-political-campaigns> accessed 15 September 2022.

14 Balazs Bodó, Natali Helberger and Claes de Vreese, 'Political micro-targeting: a Manchurian candidate or just a dark horse? Towards the next generation of political micro-targeting research' (2017) 6(4) *Internet Policy Review* <https://policyreview.info/articles/analysis/political-micro-targeting-manchurian-candidate-or-just-dark-horse> accessed 16 September 2022.

15 Sophie Boerman, Sanne Kruikemeier and Frederik Zuiderveen Borgesius, 'Online Behavioral Advertising: A Literature Review and Research Agenda' (2017) 46 *Journal of Advertising* 363. Muhammad Ali and others, 'Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes' (2019) 3 *Proceedings of the ACM on human-computer interaction* 1.

These concerns about online advertising are compounded by the fact that the online advertising ecosystem is difficult to monitor, which undermines efforts to identify, diagnose and remedy potential harms.[16] This opacity is due to personalisation: personalised advertisements are invisible to everyone except the specific users they target, hiding them from observation by outsiders.[17] As Benkler, Faris and Roberts observe, this distinguishes online advertisers from mass media advertisers, who necessarily acted 'in the public eye', thus 'suffering whatever consequences' a given message might yield outside of its target audience.[18] As a result, the online advertising ecosystem exhibits structural information asymmetries between, on one side, online platforms and advertisers, and on the other, members of the public who might hold them accountable. Researchers can potentially resort to data scraping methods, but these suffer from severe limitations and are vulnerable to interference by the platforms they target.[19] Accordingly, targeted advertising creates structural information asymmetries between advertisers and their publics.

These concerns over online political advertising took centre stage in the 'techlash', which followed the unexpected outcomes of the 2016 Brexit referendum and US presidential elections. In the UK, the Vote Leave campaign was accused of deceptive messaging, and violations of data protection law and campaign spending law in their political micro-targeting activities.[20] In the US, ad spending from Russian entities such as the Internet Research Agency raised concerns about foreign election interference. In both countries, Facebook shared selected advertising data sets in response to parliamentary investigations.[21] But these came well over a year after

16 Chester and Montgomery, 'The role of digital marketing in political campaigns' (n 13).

17 Saikat Guha, Bin Cheng and Paul Francis, 'Challenges in measuring online advertising systems' (2020). *Proceedings of the 10th Annual Conference on Internet Measurement - IMC '10* 81.

18 Yochai Benkler, Robert Faris and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press 2018), 372.

19 Bodó and others, 'Political micro-targeting: a Manchurian candidate or just a dark horse?' (n 14). Merill and Tobin, 'Facebook Moves to Block Ad Transparency Tools —Including Ours' (n 8).

20 Rob Merrick, 'Brexit: Leave "very likely" won EU referendum due to illegal overspending, says Oxford professor's evidence to High Court', *The Independent* (25 December 2019) <https://www.independent.co.uk/news/uk/politics/vote-leave-referendum-overspending-high-court-brexit-legal-challenge-void-oxford-professor-a8668771.html> accessed 22 September 2022. Jonathan Waterson, 'Obscure pro-Brexit group spends tens of thousands on Facebook ads', *The Guardian* (14 January 2019) <https://www.theguardian.com/politics/2019/jan/14/obscure-pro-brexit-group-britains-future-spends-tens-of-thousands-on-facebook-ads> accessed 20 September 2020.

21 Natasha Lomas, 'Facebook finally hands over leave campaign Brexit ads', *Techcrunch* (26 July 2018) <https://techcrunch.com/2018/07/26/facebook-finally-hands-over-leave-campaign-brexit-ads/> accessed 19 September 2022. Scott Shane, 'These are the Ads Russia Bought on Facebook in 2016', *The New York Times* (1 November 2017) <https://www.nytimes.com/2017/11/01/us/politics/russia-2016-election-facebook.html> accessed 19 September 2022.

the events actually took place—driving home the general lack of transparency and accountability in the advertising ecosystem. Similar controversies have also played out subsequent elections and referenda, such as the Irish abortion referendum of 2018 which drew an influx of foreign pro-life advertisements.[22] The actual political and electoral impact of these ad buys remains debatable.[23] But in any case, these developments drew attention to the potential for abuse in targeted advertising, and fuelled the push for more regulation and oversight in this space.

Ad archives have formed a key part of the policy response to these developments. The most prominent effort in the US is the Honest Ads Act, proposed on 19 October 2017, which would require online platforms to 'maintain, and make available for online public inspection in machine readable format, a complete record of any request to purchase on such online platform a qualified political advertisement'.[24] This bill has not yet passed (Montellaro, 2019). But only several days after its announcement, Facebook declared its plans to voluntarily build an ad archive, which would largely conform to the same requirements.[25] Google and Twitter followed suit the next year.

Since 2018, governments have started developing binding legislation on ad archives, often with resistance from platforms. Canada's Elections Modernization Act of December 2018 compels platforms to maintain public registers of political advertising sold through their service.[26] Facebook and Twitter have sought to comply with these measures, but Google instead responded by discontinuing the sale of political advertisements in this jurisdiction altogether.[27] Similarly, the State of Washington's Public Disclosure Commission attempted to regulate ad archives by requiring

22 Alex Hern, 'Facebook to block foreign spending on Irish abortion vote ads', *The Guardian* (8 May 2018) <https://www.theguardian.com/world/2018/may/08/facebook-to-block-foreign-spending-on-irish-abortion-vote-ads-referendum> accessed 24 September 2022.

23 Alan Macleod, 'Fake News, Russian Bots and Putin's Puppets'. In A. MacLeod (ed.), *Propaganda in the Information Age: Still Manufacturing Consent* (Routledge 2019). Benkler, Faris and Roberts, *Network Propaganda* (n 18).

24 The Honest Ads Act (proposed), 115th Congress S.1989, Section 8(a)(j)(1)(a)).

25 Rob Goldman, 'Update on Our Advertising Transparency and Authenticity Efforts', *Facebook Newsroom* (27 October 2017) <https://newsroom.fb.com/news/2017/10/update-on-our-advertising-transparency-and-authenticity-efforts/> accessed 19 September 2022.

26 Elections Modernization Act 2018-C-76.

27 Tom Cardoso, 'Google to ban political ads ahead of federal election, citing new transparency rules'. The Globe and Mail (March 4 2019). <https://www.theglobeandmail.com/politics/article-google-to-ban-political-ads-ahead-of-federal-election-citing-new/> accessed 15 September 2022.

advertisers publicly disclose political ads sold in the state.[28] In this case, both Google and Facebook have refused to comply with the disclosure rules and instead banned political advertising in this region.[29] Citing federal intermediary liability law, the Communications Decency Act of 1996, Facebook contended it was immune to any liability for political advertising content.[30] Some reporters also claim that Facebook has lobbied to kill the Honest Ads Act, despite publicly claiming to support regulation and implement its requirements voluntarily.[31]

Europe is also poised to regulate ad archives. In the run-up to the EU elections of May 2019, the European Commission devised the Code of Practice on Disinformation, which is not a binding law but rather a co-regulatory instrument negotiated with major tech companies including Google, Facebook, Twitter, Microsoft and Mozilla.[32] By signing the Code, these companies have committed to a range of obligations from fact-checking and academic partnerships to the creation of ad archives.[33] Furthermore, leaked documents from the European Commission show that political advertisements will receive particular attention in the upcoming reform of digital services rules.[34] Member states are also exploring the regulation of ad archives. In the UK and the Netherlands, parliamentarians have expressed support for further

28 Eli Sanders, 'Washington Public Disclosure Commission Passes Emergency Rule Clarifying That Facebook and Google Must Turn Over Political Ad Data', *The Stranger* (9 May 2019) <https://www.thestranger.com/slog/2018/05/09/26158462/washington-public-disclosure-commission-passes-emergency-rule-clarifying-that-facebook-and-google-must-turn-over-political-ad-data> accessed 19 September 2022.

29 Ibid.

30 Eli Sanders, 'Facebook Says It's Immune from Washington State Law', *The Stranger* (16 October 2018) <https://www.thestranger.com/slog/2018/10/16/33926412/facebook-says-its-immune-from-washington-state-law> accessed 19 September 2022.

31 Heather Timmons and Hannah Kozlawska, 'Facebook's quiet battle to kill the first transparency law for online political ads', *Quartz* (22 March 2018) <https://qz.com/1235363/mark-zuckerberg-and-facebooks-battle-to-kill-the-honest-ads-act/> accessed 19 September 2020.

32 The Commission describes the Code as a 'self-regulatory' instrument. However, given the Commission's involvement in its development and oversight, we consider 'co-regulatory' a more apt description. See: Kuczerawy, 'Fighting online disinformation'. More generally: Christina Angelopoulos and others, 'Study of fundamental rights limitations for online enforcement through self-regulation' (Research Report Institute for Information Law 2015). Retrieved from < pure.uva.nl/ws/files/8763808/IVIR_Study_Online_enforcement_through_self_regulation.pdf> accessed 16 September 2022.

33 EU Code of Practice on Disinformation, Section II.B.

34 Alexander Fanta and Thomas Rudl, 'Leaked document: EU Commission mulls new law to regulate online platforms', *Netzpolitik* (16 July 2019). <https://netzpolitik.org/2019/leaked-document-eu-commission-mulls-new-law-to-regulate-online-platforms/#spendenleiste> accessed 15 September 2022.

regulation in, respectively, a parliamentary resolution and a committee report.[35] France has passed a binding law requiring the public disclosure of payments received for political advertisements—if not a comprehensive regulation of ad archives *per se*.[36]

Ad archives exist alongside a number of other proposals for regulating targeted advertising. One popular measure is installing user-facing disclaimers, intended to inform audiences about e.g., the identity of the advertisers, the source of their funding, and/or the reason why they are being targeted. Another approach is to regulate funding, e.g., trough spending limits, registration requirements, or restrictions on foreign advertising. Finally, targeting technology and the use of personal data can also be regulated. Some combination of these measures is found in, *inter alia*, the US Honest Ads Act, the EU's Code of Practice, Canada's Elections Modernization Act, and France and Ireland's new election laws. The EU's General Data Protection Regulation (GDPR) is also a highly relevant instrument, since it grants users information rights, and constrains the ability for advertisers to use personal data for ad targeting purposes.[37]

Of course, present ad archive initiatives are far from uniform. Definitions of e.g., the relevant platforms, disclosure obligations and enforcement mechanisms all differ. An exhaustive comparative analysis of these differences would exceed the scope of this paper. The second half of this paper discusses how these policy initiatives differ on some of the key design issues outlined above (scoping, verifying, and targeting data), and how the major platforms have responded to their demands. First, we discuss the policy principles driving this new wave of regulation.

2.3 Normative framework: what are the policy grounds for ad archives?

Ad archive initiatives have typically been presented in terms of 'transparency and accountability', but these are notoriously vague terms. The concrete benefits of ad archives have not been discussed in much depth. To whom do ad archives create accountability, and for what? The answer is necessarily somewhat abstract, since ad archives, being publicly accessible, can be used by a variety of actors in a variety

35 Parliament of the Netherlands, 'Motion for Complete Transparency in the Buyers of Political Advertisements on Facebook' (2019) Kamerstuk 35 078 Nr 21. <https://www.parlementairemonitor.nl/9353000/1/j9vvij5epmj1ey0/vkvudd248rwa> accessed 27 September 2022. Department for Digital, Culture, Media & Sport, Online Harms White Paper (2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf> accessed 27 September 2022.

36 Loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information.

37 Tom Dobber, Ronan Ó Fathaigh and Frederik Zuiderveen Borgesius, 'The regulation of online political micro-targeting in Europe', 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1440> accessed 24 September 2022.

of accountability processes. Indeed, this diversity is arguably their strength. Other advertising transparency measures have focused on particular groups of stakeholders, such as user-facing disclaimers, third party audits or academic partnerships. Ad archives, by contrast, can enable monitoring by an unrestricted range of actors, including not only academics but also journalists, activists, government authorities and even rival advertisers—each with their own diverse capacities and motivations to hold advertising accountable. In this sense, ad archives can be seen as recreating, to some extent, the public visibility that was inherent in mass media advertising and is now obfuscated by personalisation. Broadly speaking, this public visibility can be associated with two types of accountability: (a) accountability to the law, through litigation, and, (b) accountability to public norms and values, through publicity.[38]

Ad archives can contribute to law enforcement by helping to uncover unlawful practices. Although online political advertising is not (yet) regulated as extensively as its mass media counterparts, it may still violate e.g., disclosure rules and campaign finance regulations. And, as discussed previously, new rules may soon be coming. Commercial advertising, for its part, may be subject to a range of consumer protection rules, particularly in Europe, and also to competition law, unfair commercial practice law and intellectual property law. Ad archives can allow users to proactively search for violations of these rules. Such monitoring could be done by regulators, but importantly also by third parties including commercial rivals, civil rights organisations, consumer protection organisations, and so forth. These third parties might choose to litigate independently, or simply refer the content to a competent regulator. Indeed, regulators often rely on such third party input to guide their enforcement efforts, e.g., in the form of hotlines, complaints procedures and public consultations. In most cases, litigation is likely to be straightforward and inexpensive, since most platforms operate notice and takedown procedures for the removal of unlawful advertising without the need for judicial intervention.[39] Platforms can also remove advertising based on their own community standards, even if they do not violate any national laws. In this light, ad archives can contribute to enforcement in a broad sense, including not only public advertising laws but also platforms' private standards, and relying not only on public

38 Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 *European Law Journal* 447.

39 Installing such notice and takedown processes for unlawful content is a requirement under EU law. In the US, notice and takedown procedures are only required for copyright and trademark claims, and the majority of takedown occurs on a strictly voluntary basis. In practice, much of the content removed under these regimes is assessed on the basis of platforms' voluntary standards. Daphne Keller and Paddy Leerssen, 'Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation', in: Persily N. and Tucker J (eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge University Press 2020).

authorities but on any party with the time and interest to flag prohibited content.

In addition to litigation, ad archives also facilitate publicity about advertising practices, which can serve to hold platforms accountable to public norms and values. Journalists, researchers and other civil society actors can draw on archives to research and publicise potential wrongdoings that might previously have flown under the radar. For instance, the US media has a strong tradition of analysing and fact-checking television campaign ads; ad archives could help them do similar coverage of online political micro-targeting. Such publicity may encourage platforms and/or advertisers to self-correct and improve their advertising standards, by raising the threat of reputational harm. And failing such a private ordering response, publicity can also provide an impetus for new government interventions. In these ways, ad archives can contribute not only to the enforcement of existing laws, but also to informed public deliberation, and thus to the articulation and enforcement of public norms and values.[40] Such publicity effects may be especially important in the field of online political advertising, since, as discussed, this space remains largely unregulated under existing laws, and raises many novel policy questions for public deliberation.

In each case, it is important to note the factor of deterrence: the mere *threat* of publicity or litigation may already serve to discipline unlawful or controversial practices. Even for actors who have not yet faced any concrete litigation or bad publicity, ad archives could theoretically have a disciplinary effect. In this sense, a parallel can be drawn with the concept of the Panopticon, as theorised in surveillance studies literature; subjects are disciplined not merely through the fact of observation, but more importantly through the pervasive possibility of observation.[41] Put differently, Richard Mulgan describes this as the *potentiality* of accountability; the possibility that one 'may be called to account for anything at any time'.[42] Or, as the saying goes: The value in the sword of Damocles is not that it drops, but that it hangs.[43]

Of course, these accountability processes depend on many other factors besides transparency alone. Most importantly, ad archives depend on a capable and motivated user base of litigators (for law enforcement effects) and civil society watchdogs (for publicity effects). For publicity effects, these watchdogs must also be sufficiently

40 José van Dijck, Thomas Poell and Martijn de Waal, *The platform society: Public values in a connective world* (Oxford University Press 2018).

41 Michel Foucault, *Discipline and Punish: The Birth of the Prison* (1977) Pantheon Books. David Lyon, *Theorizing Surveillance: The Panopticon and Beyond* (Willan Publishing 2006).

42 Richard Mulgan, 'Accountability: An Ever-Expanding Concept?' (2000) 78 *Public Administration* 555.

43 *Arnett v Kennedy* (1974) 416 U.S. 134 (Supreme Court of the United States, 1974)

influential to create meaningful reputational or political risks for platforms.[44] These conditions can certainly not be assumed; which researchers are up to the task of overseeing this complex field, and holding its powerful players to account? This may call for renewed investment in our public watchdogs, including authorised regulators as well as civil society. Ad archives might be a powerful tool, but they rely on competent users.

Finally, of course, the above analysis also assumes that ad archives are designed effectively, so as to offer meaningful forms of transparency. As we discuss in the following section, present implementations leave much to be desired.

3. Pitfalls: key challenges for ad archive architecture

Having made the basic policy case for the creation of ad archives, we now discuss several criticisms of current ad archive practice. Firstly, we discuss the issue of scoping: which ads are included in the archive? Second, verifying: how do ad archives counteract inauthentic behaviour from advertisers and users? Third, targeting: how do ad archives document ad targeting practices? Each of these issues can create serious drawbacks to the research utility of ad archives, and deserve further scrutiny in future governance debates.

Ad archive architecture is very much a moving target, so we emphasise that our descriptions represent a mere snapshot. Circumstances may have changed significantly since our time of writing. Accordingly, the following is not intended as an exhaustive list of possible criticisms, but rather as a basic assessment framework for some of the most controversial issues. For instance, one important criticism of ad archives which we do *not* consider in detail is the need for automated access through application programming interfaces (APIs). When ad archive data is exclusively available through browser-based interfaces, this can make it relatively time-consuming to perform large-scale data collection. To enable in-depth research, it is clear that ad archives must enable such automated access. Until recently, Facebook did not offer public API access to their ad archive data.[45] And once the API was made publicly accessible, it quickly appeared to be so riddled with bugs as to be almost unusable.[46] As noted by

44 Christopher Parsons, 'The (In)effectiveness of Voluntarily Produced Transparency Reports' (2019) *58 Business & Society* 103.

45 Satwik Shukla, 'A Better Way to Learn About Ads on Facebook', *Facebook Newsroom* (28 March 2019) <https://newsroom.fb.com/news/2019/03/a-better-way-to-learn-about-ads/> accessed 19 September 2020.

46 Matthew Rosenberg, 'Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised', *The New York Times* (25 July 2019) <https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html> accessed 19 September 2022.

Laura Edelson, these API design issues are not novel or intractable from a technical perspective, but eminently 'fixable', and thus reflect sub-standard implementation on the part of Facebook.[47] In response, Mozilla, together with a coalition of academics, has drafted a list of design principles for ad archive APIs.[48] Such public, automated access can be seen as a baseline condition for effective ad archive policy. What then remains are questions about the *contents* of the archive, which include scoping, verifying and targeting.

3.1 Scoping: what ads are included in the archive?

A key design question for ad archives is that of scope: what ads are included in the archive? First, we discuss the concept of 'political' advertising, which is the central scoping device in most existing initiatives and has led to many implementation challenges. Second, we discuss the attempts to exempt news reporting from political ad archives.

'Political' ad archives: electoral ads v. issue ads v. all ads?

Ad archive initiatives, both self-regulatory and governmental, have emphasised 'political' advertising rather than commercial advertising. However, their precise interpretations of this concept differ significantly. Below we discuss these differing approaches and relevant policy trade-offs.

The main dividing line in existing political ad archives is between issue ads and electoral ads (or 'campaign ads'). 'Election ads' explicitly reference an election or electoral candidate, whereas 'issue ads' reference a topic of national importance. Google focuses exclusively on election ads, whereas Facebook and Twitter also include issue ads in certain jurisdictions, and even non-political ads. Most public policy instruments also focus on issue ads, including the US Honest Ads Act and the EU Code of Practice. There is good reason to include issue ads, since they have been central to recent controversies. During the 2016 US election, for instance, foreign actors such as the Russian-controlled Internet Research Agency advertised on divisive issues such as racial politics, sexual politics, terrorism, and immigration, in an apparent attempt to influence the election.[49] An approach which focuses on election ads would fail to address such practices.

47 Ibid.

48 Mozilla, 'Data Collection Log — EU Ad Transparency Report' (Research Report Mozilla 2019) <https://adtransparency.mozilla.org/eu/log/> accessed 19 September 2022.

49 Philip Howard and others, 'The IRA, Social Media and Political Polarization in the United States, 2012–2018' (Research Report Oxford Computational Propaganda Project 2018) < https://www.oii.ox.ac.uk/news-events/reports/the-ira-social-media-and-political-polarization-in-the-united-states-2012-2018/> accessed 15 September 2022.

However, the drawback of 'issue ads' as a scoping device, is that the concept of a political 'issue' is broad and subjective, and makes it difficult for archive operators to develop actionable definitions and enforce these in practice. Google, in its implementation reports for the EU's Code of Practice, reported difficulties in developing a workable definition of a 'political issue'.[50] The European Commission later lamented that 'Google and Twitter have not yet reported further progress on their policies towards issue-based advertising' (European Commission, 2019). In Canada, where the Election Act also requires the disclosure of issue-based ads, Google has claimed that they are simply unable to comply with disclosure requirements.[51] These difficulties might explain why the company announced plans, as discussed previously, to ban political advertising entirely for Canadian audiences during election periods.

Yet these attempts to ban political advertising, as an alternative to disclosure, beg the question whether platforms can actually enforce such a ban. After all, the platforms themselves admit they struggle to identify political ads in the first place. Simply declaring that political ads are prohibited will not guarantee that advertisers observe the ban and refrain from submitting political content. Could platforms then still be liable for a failure to disclose? Here, a tension emerges between ad archive regulation and intermediary liability laws, which typically immunise platforms for (advertising) content supplied by their users. Canada, Europe and the US all have such laws, although their precise scope and wording differ. Indeed, Facebook has argued that it is immunised against Washington State's disclosure rules based on US federal intermediary liability law—the Communications Decency Act of 1996.[52] Similarly, the EU's intermediary safe harbours, which prohibit 'proactive monitoring obligations' imposed on platforms.[53] Such complex interactions with intermediary liability law should be taken into account in ongoing reforms.

Compared to Google, Facebook is relatively advanced in its documentation of issue ads. But that company too has faced extensive criticism for its approach. The company employs approximately 3,000-4,000 people in reviewing ads related to politics or issues, using 'a combination of artificial intelligence (AI) and human review', and

50 Google, Implementation Report for EU Code of Practice on Disinformation (2019) <https://ec.europa.eu/information_society/newsroom/image/document/2019-5/google_-_ec_action_plan_reporting_CF162236-E8FB-725E-C0A3D2D6CCFE678A_56994.pdf> accessed 26 September 2022.

51 Tom Cardoso, 'Google to ban political ads ahead of federal election, citing new transparency rules' (n 27).

52 Eli Sanders, 'Facebook Says It's Immune from Washington State Law', The Stranger (16 October 2018) <https://www.thestranger.com/slog/2018/10/16/33926412/facebook-says-its-immune-from-washington-state-law> accessed 19 September 2022.

53 E-Commerce Directive 2000/31/EC, Article 15.

is estimated to process upwards of a million ad buyers per week in the US alone.[54] Facebook's website offers a list of concrete topics which they consider 'political issues of national importance', tailored to the relevant jurisdiction. The US list of political issues contains 20 entries, including relatively specific ones such as 'abortion' and 'immigration', but also relatively broad and ambiguous ones such as 'economy' and 'values'.[55] The EU list contains only six entries so far, including 'immigration', 'political values' and 'economy'.[56]

Despite these efforts, research suggests that Facebook's identification of political issue ads is error-prone. Research from Princeton and Bloomberg showed that a wide range of commercial ads are at risk of being mislabeled as political, including advertisements for e.g., national parks, veteran's day celebrations, and commercial products that included the words 'bush' or 'clinton'.[57] Conversely, data scraping research by ProPublica shows that Facebook failed to identify political issue ads on such topics as civil rights, gun rights, electoral reform, anti-corruption, and health care policy.[58] These challenges are likely to exacerbate as platforms expand their efforts beyond the United States to regions such as Africa and Europe, which contain far greater political and linguistic diversity and fragmentation. Accordingly, further research is needed to determine whether the focus on issue ads in ad archives is appropriate. It may appear in future that platforms are able to refine their processes and identify issue ads with adequate accuracy and consistency. But given the major scaling challenges, the focus on issue ads may well turn out to be impracticable.

In light of the difficulties with identifying 'issue ads', one possible alternative would be to simply include all ads *without* an apparent commercial objective. In other words, a definition *a contrario*. This approach could capture the bulk of political advertising, and would avoid the difficulties of identifying and defining specific political 'issues'. Such an approach would likely be more scalable and consistent than the current model, although this might come at the cost of increased false positives (i.e., a greater overinclusion of irrelevant, non-political ads in the archive).

---

54 Matias, Hounsel and Hopkins, 'We tested Facebook's ad screeners and some were too strict' (n 2).

55 Facebook, 'Ads about social issues, elections or politics', *Facebook Business Help Center* (n.d.) <https://www.facebook.com/business/help/208949576550051> accessed 2022

56 Ibid.

57 Austin Hounsel and others, 'Estimating Publication Rates of Non-Election Ads by Facebook and Google'. Github (1 November 2019). <https://github.com/citp/mistaken-ad-enforcement/blob/master/estimating-publication-rates-of-non-election-ads.pdf> accessed 19 September 2022.). Sarah Frier, 'Facebook's Political Rule Blocks Ads for Bush's Beans, Singers Named Clinton' (2 July 2018) *Bloomberg* <https://www.bloomberg.com/news/articles/2018-07-02/facebook-s-algorithm-blocks-ads-for-bush-s-beans-singers-named-clinton> accessed 15 September 2022.

58 Merrill and Tobin, 'Facebook Moves to Block Ad Transparency Tools —Including Ours' (n 8).

Another improvement could be to publish *all* advertisements in a comprehensive archive, regardless of their political or commercial content.[59] This would help third parties to independently evaluate platforms' flagging processes for political ads, and furthermore to research political advertising according to their *own* preferred definitions of the 'political'. This is what Twitter does in its Ad Transparency Center: the company still takes steps to identify and flag political advertisers (at least in the US), but users have access to all other ads as well.[60] However, only political ads are accompanied by detailed metadata, such as ad spend, view count, targeting criteria, et cetera. Facebook, in an update from 29 March 2019, also started integrating commercial ads into its database.[61] Like Twitter, however, these ads are not given the same detailed treatment as political ads. In this light, Twitter and Facebook appear to be moving towards a *tiered* approach, with relatively more detail on a subset of political ads, and relatively less detail on all other ads.

Of course, a more fundamental advantage of comprehensive publication ads is that it extends the benefits of ad archives to *commercial* advertising. Commercial advertising has not been the primary focus of ad archive governance debates thus far, but here too ad archives could be highly beneficial. A growing body of evidence indicates that online commercial ad delivery raises a host of legal and ethical concerns, including discrimination and manipulation.[62] Furthermore, online advertising is also subject to a range of consumer protection laws, including child protection rules and prohibitions on unfair and deceptive practices. With comprehensive publication, ad archives could contribute to research and reporting on such issues, especially if platforms abandon their tiered approach and start publishing more detailed metadata for these ads.

Platforms may not be inclined to implement comprehensive ad archives since, as discussed, their commercial incentives may run counter to greater transparency. But from a public policy perspective, there appear to be no obvious drawbacks to comprehensive publication, at least as a default rule. If there are indeed grounds to

59 Philip Howard, 'A Way to Detect the Next Russian Misinformation Campaign', *The New York Times* (27 March 2019) <https://www.nytimes.com/2019/03/27/opinion/russia-elections-facebook.html?module=inline> accessed 19 September 2022.

60 Twitter, Progress Report for the EU Code of Practice on Disinformation (2019) <http://ec.europa.eu/information_society/newsroom/image/document/2019-5/twitter_progress_report_on_code_of_practice_on_disinformation_CF162219-992A-B56C-06126A9E7612E13D_56993.pdf> accessed 26 September 2022.

61 Satwik Shukla, 'A Better Way to Learn About Ads on Facebook', *Facebook Newsroom* (28 March 2019) <https://newsroom.fb.com/news/2019/03/a-better-way-to-learn-about-ads/> accessed 19 September 2020.

62 Ali and others, 'Discrimination through optimization' (n 15). Boerman, Kruikemeier and Zuiderveen Borgesius, 'Online Behavioral Advertising' (n 15).

shield certain types of ads from public archives—though we see none as of yet—such cases could also be addressed through exemption procedures. The idea of comprehensive ad archives therefore warrants serious consideration and further research, since it promises to benefit the governance of both commercial *and* political advertising.

Exemptions for news reporting

Some ad archive regimes offer exemptions for news publishers and other media actors. News publishers commonly use platform advertising services to promote their content, and when this content touches on political issues it can therefore qualify as an issue ad. Facebook decided to exempt news publishers from their ad archive in 2018, following extensive criticism from US press industry trade associations, who penned several open letters criticising their inclusion in ad archives. They argued that '[t]reatment of quality news as political, even in the context of marketing, is deeply problematic' and that the ad archive 'dangerously blurs the lines between real reporting and propaganda'.[63] Similar exemptions can now also be found in Canada's Elections Modernization Act and in the EU Code of Practice.[64] However, the policy grounds for these exemptions are not particularly persuasive. There is little evidence to suggest, or reason to assume, that inclusion in ad archives would meaningfully constrain the press in its freedom of expression. Indeed, ad archive data about media organisations is highly significant, since the media are directly implicated in concerns about misinformation and electoral manipulation.[65] Excluding the media's ad spending is therefore a missed opportunity without a clear justification.

3.2 Verifying: how do archives account for inauthentic behaviour?

Another pitfall for ad archives is verifying their data in the face of fraud and other inauthentic behaviours. One key challenge is documenting ad buyers' identities. Another is the circumvention of ad archive regimes by 'astroturf', sock puppets and other forms of native advertising. More generally, engagement and audience statistics may be inaccurate due to bots, click fraud and other sources of noise. As we discuss below, these pitfalls should serve as a caution to ad archive researchers, and as a point of attention for platforms and their regulators.

63 Alfredo Carbajal and others, 'Open Letter to Marck Zuckerberg on Alternative Solutions for Politics Tagging' (2018) News Media Alliance. <https://www.newsmediaalliance.org/wp-content/uploads/2018/06/vR_Alternative-Facebook-Politics-Tagging-Solutions-FINAL.pdf> accessed 15 September 2022. David Chavern, 'Open Letter to Mr. Zuckerberg', News Media Alliance (18 May 2018). Retrieved from <http://www.newsmediaalliance.org/wp-content/uploads/2018/05/FB-Political-Ads-Letter-FINAL.pdf> accessed 15 September 2022.

64 Rob Leathern, 'Updates to our ad transparency and authorisation efforts' (29 November 2018). <https://www.facebook.com/facebookmedia/blog/updates-to-our-ads-transparency-and-authorisation-efforts> accessed 15 September 2019.

65 Benkler, Faris and Roberts, *Network Propaganda* (n 18).

Facebook's archive in particular has been criticised for failing to reliably identify ad buyers.[66] Until recently, Facebook did not verify the names that advertisers submitted for their 'paid for by' disclaimer. This enabled obfuscation by advertisers seeking to hide their identity.[67] For instance, ProPublica uncovered 12 different political ad campaigns that had been bought in the name of non-existent non-profits, and in fact originated from industry trade organisations such as the American Fuel & Petrochemical Manufacturers.[68] Vice Magazine even received authorisation from Facebook to publish advertisements in the name of sitting US senators.[69] More recently, Facebook has therefore started demanding proof of ad buyer identity in several jurisdictions, such as photo ID and notarised forms.[70] Twitter and Google enforce similar rules.[71] The Canadian Elections Modernization Act now codifies these safeguards by requiring platforms to verify and publish ad buyers' real names.[72]

Such identity checks are only a first step in identifying ad buyers, however. Ad buyers wishing to hide their identity can still attempt to purchase ads through proxies or intermediaries. In theory, platforms could be required to perform even more rigorous background checks or audits so as to determine their ultimate revenue sources. But there may be limits to what can and should be expected of platforms in this regard. Here, ad archive governance intersects with broader questions of campaign finance regulation and the role of 'dark money' in politics. These issues have historically been tackled through national regulation, including standardised registration mechanisms for political advertisers, but many of these regimes currently do not address online advertising. Platforms' self-regulatory measures, though useful as a first step, cannot make up for the lack of public regulation in this space.[73] Even Facebook CEO Mark Zuckerberg has called

66 Edelson and others, 'An Analysis of United States Online Political Advertising Transparency' (n 2).

67 Albright, 'Facebook and the 2018 Midterms' (n 2). Andringa, 'Interactive: See Political Ads Targeted to You on Facebook' (n 2). Lapowsky, 'Obscure Concealed-Carry Group Spent Millions on Facebook Political Ads' (n 2). O'Sullivan, 'What an anti-Ted Cruz meme page says about Facebook's political ad policy' (n 2). Waterson, 'Obscure pro-Brexit group spends tens of thousands on Facebook ads' (n 2).

68 Merrill, 'How Big Oil Dodges Facebook's New Ad Transparency Rules' (n 8).

69 William Turton, 'We posed as 100 senators to run ads on Facebook. Facebook approved all of them', *VICE News* (30 October 2018) <https://news.vice.com/en_ca/article/xw9n3q/we-posed-as-100-senators-to-run-ads-on-facebook-facebook-approved-all-of-them> accessed 19 September 2022.

70 Facebook, 'Ads about social issues, elections or politics', *Facebook Business Help Center* (2019 n.d.) <https://www.facebook.com/business/help/208949576550051> accessed 2022.

71 Google, 'Verification for election advertising in the European Union' (n.d.) <https://support.google.com/adspolicy/answer/9211218> accessed 26 September 2022. Twitter, 'How to get certified as a political advertiser', *Twitter Business* (2019) <https://business.twitter.com/en/help/ads-policies/restricted-content-policies/political-content/how-to-get-certified.html>.

72 Elections Modernization Act 2018-C-76 (Canada), Section 325.1(3)(b) and 325.2A.

73 Livingstone and others, 'Tackling the Information Crisis' (n 7).

for regulation here, arguing in a recent op-ed that '[o]ur systems would be more effective if regulation created common standards for verifying political actors'.[74]

Another weak spot for ad archives is that they fail to capture native advertising practices: advertising which is not conducted through social media platforms' designated advertising services, but rather through their organic content channels. Such 'astroturfing' strategies, as they are also known, have seen widespread deployment in both commercial and political contexts, from Wal-Mart and Monsanto campaigns to Russian 'troll farms' to presidential Super PACs.[75] Ad archives do not capture this behaviour, and indeed their very presence could further *encourage* astroturfing, as a form of regulatory arbitrage. Benkler, Faris, and Roberts suggest that ad archive regulation should address this issue by imposing an independent duty on advertisers to disclose any 'paid coordinated campaigns' to the platform.[76] One example from practice is the Republic of Ireland's Online Advertising and Social Media Bill of 2017, which would hold ad buyers liable for providing inaccurate information to ad sellers, and also prohibit the use of bots which 'cause multiple online presences directed towards a political end to present as an individual account or profile on an online platform'.,[77] Enforcing such rules will remain challenging, however, since astroturfing is difficult to identify and often performed by bad actors with little or no interest in complying with the law.[78]

For ads that are actually included in the archive, inauthentic behaviour can also distort associated metadata such as traffic data. Engagement metrics, including audience demographic data, can be significantly disturbed by click fraud or bot traffic.[79] Platforms typically expend extensive resources to combat inauthentic behaviour, and this appears to be a game of cat-and-mouse without definitive solutions. In light of these challenges,

74 Mark Zuckerberg, 'The Internet needs new rules. Let's start in these four areas', *The Washington Post* (30 March 2019) <https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html> accessed 20 September 2020.

75 Ben Collins, 'Hillary PAC Spends $1 Million to 'Correct' Commenters on Reddit and Facebook' (21 April 2016). <https://www.thedailybeast.com/articles/2016/04/21/hillary-pac-spends-1-million-to-correct-commenters-on-reddit-and-facebook> accessed 15 September 2022. Howard and others, 'The IRA, Social Media and Political Polarization in the United States' (n 59). Mark Leiser, 'AstroTurfing, 'CyberTurfing' and other online persuasion campaigns' (2016) 7(1) *European Journal of Law and Technology* <https://ejlt.org/index.php/ejlt/article/view/501> accessed 19 September 2022.

76 Benkler, Faris and Roberts, *Network Propaganda* (n 18).

77 Online Advertising and Social Media (Transparency) Bill of 2017.

78 Leiser, '"Astroturfing", "CyberTurfing" and other online persuasion campaigns' (n 75).

79 Benjamin Edelman, 'Pitfalls and Fraud In Online Advertising Metrics: What Makes Advertisers Vulnerable to Cheaters, And How They Can Protect Themselves' (2014) 54 *Journal of Advertising Research* 127. Gian Fulgoni, 'Fraud in Digital Advertising: A Multibillion-Dollar Black Hole: How Marketers Can Minimize Losses Caused by Bogus Web Traffic' (2016) 56 *Journal of Advertising Research* 122.

researchers should maintain a healthy scepticism when dealing with ad archive data and, where necessary, continue to corroborate ad archive findings with alternative sources and research methods.[80]

The above is not to say that *all* information supplied by ad buyers should be verified. There may still be an added value in enabling voluntary, unverified disclosures by ad buyers in archives. Facebook, for instance, gives advertisers the option to include 'Information From the Advertiser' in the archive. Such features can enable good faith advertisers to further support accountability processes, e.g., by adding further context or supplying contact information. It is essential, however, that such unverified submissions are recognisably earmarked as such. Ad archive operators should clearly describe which data is verified, and how, so that users can treat their data with the appropriate degree of scepticism.

3.3 Targeting: how is ad targeting documented?

Another key criticism of ad archives is that they are not detailed enough, particularly in their documentation of ad targeting practices. Micro-targeting technology, as discussed previously, is the source of many public policy concerns for both political and commercial advertising, including discrimination, deception, and privacy harms. These threats are relatively new, and are both undocumented and unregulated in many jurisdictions—particularly as regards political advertising.[81] Regrettably, ad archives currently fail to illuminate these practices in any meaningful depth.

At the time of writing, the major ad archives differ significantly in their approach to targeting data. Google's archive indicates whether the following targeting criteria have been selected by the ad buyer: age, location, and gender. It also lists the top five Google keywords selected by the advertiser. Facebook's Ad Library, by contrast, does not disclose what targeting criteria have been selected, but instead shows a demographic breakdown of the actual audience that saw the message—also in terms of age, location and gender. Twitter offers *both* audience statistics and targeting criteria, and covers not only the targeting criteria of age, location, and gender, but also their preferred language. These data vary in granularity. For instance, Google's archive lists six different age brackets between the ages of 18 and 65+, whereas Twitter lists 34. For anyone familiar with the complexities of online behavioural targeting, it is apparent that these datasets leave many important questions unanswered. These platforms offer far more refined methods for ad targeting and performance tracking than the basic features described above.

80 Eveline Vlassenroot and others, 'Web archives as a data resource for digital scholars' (2019) 1 *International Journal of Digital Humanities* 85.

81 Bodó, Helberger and De Vreese, 'Political micro-targeting: a Manchurian candidate or just a dark horse?' (n 14).

For better insights into ad targeting, one helpful rule of thumb would be to insist that ad archives should include an equivalent level of information as is offered to the actual ad buyer—both in terms of targeting criteria and in terms of actual audience demographics.[82] For some targeting technologies, full disclosure of targeting practices might raise user privacy concerns. For instance, Facebook's Custom Audience feature enables advertisers to target users by supplying their own contact information, such as email addresses or telephone numbers. Insisting on full disclosure of targeting criteria for these custom audiences would lead to the public disclosure of sensitive personal data.[83] Anonymisation of these data may not always be reliable.[84] In these cases, however, Facebook could at a minimum still disclose any additional targeting criteria selected by the ad buyer in order to refine this custom audience. Furthermore, ad performance data, rather than ad targeting data, can also provide some insight into targeting without jeopardising the custom audience's privacy.[85] Other platforms' advertising technologies might raise comparable privacy concerns, demanding a case-by-case assessment of relevant tradeoffs. These exceptions and hard cases notwithstanding, however, there are no clear objections (either technical or political) that should prevent platforms from publicly disclosing the targeting methods selected by their advertisers.

In light of such complexities, designing appropriate disclosures will likely require ongoing dialogue between archive operators, archive users and policymakers. The first contours of such a debate can already be found in the aforementioned work of Edelson et al., Rieke and Bogen, and Mozilla, who have done valuable work in researching and critiquing early versions of Google, Twitter and Facebook's data sets.[86] For the time being, researchers may also choose to combine ad archive data with other sources, such as Facebook's Social Science One initiative, or GDPR data access rights, in order to obtain a more detailed understanding of targeting practices.[87] For instance, Ghosh et al. supplemented ad archive research with data scraped with ProPublica's research tool, which gave insights into ad

82 Mozilla, 'Facebook and Google: This is What an Effective Ad Archive API Looks Like', *The Mozilla Blog* (2019, March 27) <https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like> accessed 19 September 2022.

83 Rieke and Bogen, 'Leveling the Platform' (n 2).

84 Paul Ohm (2010), 'Broken Promises of Privacy: Responding To The Surprising Failure of Anonymization' (2010) 57 *UCLA Law Review* 1701.

85 Rieke and Bogen, 'Leveling the Platform' (n 2).

86 Edelson and others, 'An Analysis of United States Online Political Advertising Transparency' (n 2). Mozilla, 'Facebook and Google: This is What an Effective Ad Archive API Looks Like' (n 2).

87 Jef Ausloos, 'GDPR Transparency as a Research Method' (2019) SSRN Working paper. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3465680> accessed 16 September 2022. Christian Sandvig and others, 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms' (2014), Paper presented to Data and Discrimination: Converting Critical Concerns into Productive Inquiry <https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf> accessed 26 September 2022.

targeting that were not offered through the ad archive.[88] Along these lines, ad archives can help to realise Pasquale's model of 'qualified transparency', which combines general public disclosures with more limited, specialist inquiries.[89]

4. Conclusion

This paper has given an overview of a new and rapidly developing topic in online advertising governance: political ad archives. Here we summarise our key findings, and close with suggestions for future research in both law and communications science.

Ad archives can be a novel and potentially powerful governance tool for online political advertising. If designed properly, ad archives can enable monitoring by a wide range of stakeholders, each with diverse capacities and interests in holding advertisers accountable. In general, ad archives can not only improve accountability to applicable laws, but also to public opinion, by introducing publicity and thus commercial and political risk into previously invisible advertisements.

Public oversight will likely be necessary to realise these benefits, since platforms ostensibly lack the incentives to voluntarily optimise their ad archives for transparency and accountability. Indeed, our analysis here has already identified several major shortcomings in present ad archive policies: scoping, verifying, and targeting. To realise the full potential of ad archives, these issues will require further research, critique, and likely regulation. Our review suggests that major advances can already be made by comprehensively publishing *all* advertisements, regardless of whether they have been flagged as political; revoking any exemptions for media organisations; requiring basic verification of ad buyers' identities; documenting how ad archive data is verified; and disclosing all targeting methods selected by the ad buyer (insofar as possible without publishing personal data).

Looking forward, ad archives present a fruitful research area for both legal and communication sciences scholars. For legal scholars, the flurry of law making around political advertising in general, and transparency in particular, raises important questions about regulatory design (in terms of how relevant actors and duties are

88 Avijit Ghosh, Giridhari Venkatadri and Alan Mislove, 'Analyzing Political Advertisers' Use of Facebook's Targeting Features' (2019) *IEEE workshop on technology and consumer protection* <https://www.ieee-security.org/TC/SPW2019/ConPro/papers/ghosh-conpro19.pdf> accessed 19 September 2022.

89 Frank Pasquale, *The Black Box Society: The secret algorithms that control money and information* (Harvard University Press 2015).

defined, oversight and enforcement mechanisms, etc.). In future, ad archives also deserve consideration in commercial advertising governance, in such areas as consumer protection, child protection, or anti-discrimination.

The emergence of ad archives also has important implications for communications science. Firstly, ad archives could become an important *resource* of data for communications research, offering a range of data that would previously have been difficult or impossible to obtain. Although our paper has identified several shortcomings in this data, they might nonetheless provide a meaningful starting point to observe platforms' political advertising. Secondly, ad archives are an interesting *object* of communications science research, in terms of how they are used by relevant stakeholders, and how this impacts advertising and communications practice. Further research along these lines will certainly be necessary to better understand ad archives, and to make them reach their full potential.

3

CHAPTER 4

News from the ad archive: How journalists use the Facebook Ad Library to hold online advertising accountable[1]

1 Originally published as: Paddy Leerssen, Tom Dobber, Natali Helberger and Claes de Vreese, 'News from the ad archive: how journalists use the Facebook Ad Library to hold online advertising accountable' (2021) *Information, Communication and Society* <https://doi.org/10.1080/1369118X.2021.2009002>. Two appendices for this article are also include at the end of this dissertation. Appendix I contains our content analysis protocol. Appendix II describes supplemental keyword testing conducted as part of the peer review process.

Abstract

The Facebook Ad Library promises to improve transparency and accountability in online advertising by rendering personalised campaigns visible to the public. This article investigates whether and how journalists have made use of this tool in their reporting. Our content analysis of print journalism reveals several different use cases, from high-level reporting on political campaigns to uncovering specific wrongdoings such as disinformation, hate speech, and astroturfing. However, our interviews with journalists who use the Ad Library show that they remain highly critical of this tool and its manifold limitations. We argue that these findings offer empirical grounding for the public regulation of ad archives, since they underscore both the public interest in advertising disclosures as well as the growing reliance of journalists on voluntary and incomplete access frameworks controlled by the very platforms they aim to scrutinise.

1. Introduction

Since 2018, major advertising platforms have started to publish ad archives: public databases documenting advertisements sold on their services. These reforms respond to mounting concerns about the lack of transparency and accountability in this industry. Several governments are now poised to regulate ad archives by law, as in Canada's Elections Modernization Act and the EU's proposed Digital Services Act. Lively academic debate has ensued as to the merits of ad archives, including a growing body of evidence pointing to the shortcomings of existing self-regulatory efforts.

Now, over three years since the launch of the Facebook Ad Library, the earliest and most expansive platform ad archive, this article offers a first attempt to map its impact in practice. It asks not what usage this tool *could* enable but rather what usage it *has* enabled. In particular, this paper examines journalists as a key user group, central to public accountability processes. It inquires whether and how journalists have made use of Facebook Ad Library in their reporting, and whether these practices contribute to accountability in online advertising.

This paper proceeds as follows: Section II describes the Ad Library and its features, the policy concerns that drove its creation, and its relevance to watchdog journalism. Section III provides a content analysis of ad archive journalism: through an inductive, quantitative pilot study, we generate a typology of different journalistic usages of the Ad Library. On this basis we perform a quantitative analysis of print journalism sampled from the LexisNexis database in order to appraise the composition and scale of this phenomenon. Section III then describes interviews with relevant journalists, which review their experiences with and attitudes towards the Ad Library. In light of these findings, Section IV assesses the Ad Library's contribution to transparency and accountability in platform governance.

2. Background

2.1 The Ad Library and its features

The Ad Library documents a selection of ads that appeared on Facebook.[2] It lists the ad content as well as metadata such as buyer name, amount spent, and demographic

2 Readers should note that the Ad Library's policies and affordances change frequently, and may have changed since our time of writing. Our description is based on public Facebook policies as of July 2021.

reach by region, age and gender.[3] The Ad Library is available through a browser interface as well as an automated programming interface (API). Currently, the Ad Library focuses primarily on 'ads about social issues, elections and politics'.[4] Advertisers seeking to publish ads in this category must apply for prior authorisation, and Facebook enforces this rule through human and automated monitoring. Facebook maintains lists of political 'issues' for several jurisdictions in order to operationalise their classifications.[5] Other (commercial) ads receive a lower level of transparency: they are only visible as long as they are active on the platform, with restricted search functionalities and less metadata.[6]

The Ad Library has been criticised extensively for its faulty design and implementation.[7] To name some of the most significant shortcomings: the Ad Library's demographic data does not disclose the targeting mechanisms involved; its audience and spend data are insufficiently granular; the browser interface and API are restrictive and unreliable; the focus on political and issue ads is restrictive, and its definitions are ambiguous and subjective; the identification of these ads in practice has proven inconsistent, leading to both false positives and false negatives; and data are not standardised across different platforms. Analysis by Laura Edelson

3 Facebook, 'Ads about social issues, elections or politics', *Facebook Business Help Center* (n.d.) <https://www.facebook.com/business/help/208949576550051> accessed 2022 accessed 15 September 2022.

4 Ibid.

5 Ibid.

6 Ibid.

7 Chapter 3 above (Paddy Leerssen, Jef Ausloos, Brahim Zarouali, Natali Helberger and Claes de Vreese, 'Platform ad archives: promises and pitfalls' (2019) 8(4) Internet Policy Review <https://doi.org/10.14763/2019.4.1421>) Aaron Rieke and Miranda Bogen, Leveling the Platform: Real Transparency for Paid Messages on Facebook. (Research Report UpTurn 2018) <https://www.upturn.org/static/reports/2018/facebook-ads/files/Upturn-Facebook-Ads-2018-05-08.pdf> accessed 19 September 2022. Mozilla, 'Facebook and Google: This is What an Effective Ad Archive API Looks Like', The Mozilla Blog (27 March 2019) <https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like> accessed 19 September 2022. Laura Edelson and others, 'An Analysis of United States Online Political Advertising Transparency' (2019) *ArXiv [Cs]* <http://arxiv.org/abs/1902.04385 accessed 15 September 2022> accessed 19 September 2022. Laura Edelson, Thomas Lauinger and Damian McCoy, 'A Security analysis of the Facebook Ad library' (2020) *IEEE Symposium on Security and Privacy* 661. Márcio Silva and others, 'Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook' (2020) *WWW '20: Proceedings of the Web Conference 2020* 224. Austin Hounsel and others, 'Estimating Publication Rates of Non-Election Ads by Facebook and Google'. Github (1 November 2019). <https://github.com/citp/mistaken-ad-enforcement/blob/master/estimating-publication-rates-of-non-election-ads.pdf> accessed 19 September 2022.). Jennifer Grygiel and Weston Sager, 'Unmasking Uncle Sam: A Legal test for identifying State media' (2020) 11 *UC Irvine Law Review* 383. Daniel Kreiss and Bridget Barrett, 'Democratic tradeoffs: Platforms and political advertising' (2020) 16 *Ohio State Technology Law Journal* 493.

et al. also showed that ads had been retroactively removed from the Ad Library, calling into question the reliability of its archival function.[8]

2.2 Background and rationale

Ad archives emerged as a response to mounting criticism of online advertising following the U.S. presidential election and U.K. Brexit campaign of 2016. Much of this criticism is closely connected to the personalised distribution of microtargeted ads, and the resulting lack of a public record. A personalised ad is in principle only visible to the specific audience members it targets, and leaves no trace after its distribution. In the legacy media, by contrast, ads are public in the sense that they are equally accessible to all audience members, in addition to commonly being preserved by institutions such as newspaper and television archives.[9] The non-public and ephemeral nature of online advertising makes it more akin to direct marketing via email or telephone, which has raised comparable policy concerns around transparency and accountability.[10]

The policy concerns related to transparency of online advertising are several. First, political microtargeting might undermine electoral accountability, by allowing campaigners to signal different campaign promises to different constituencies.[11] This 'fragmentation of the marketplace of ideas' is also seen to undermine the capacity for public deliberation, since political actors can no longer observe and respond to the microtargeted ads of their rivals.[12] As a result, the capacity for 'dark advertising' may also engender false and inflammatory messaging, by foreclosing the ability of rival campaigners, media actors and other third parties to rebut, critique or otherwise sanction such transgressions. Similarly, dark ads provide cover for 'dark money' advertising funded by special interests and foreign governments.[13] A related concern is that targeting leads to algorithmic discrimination, which may exclude people from valuable content, and, conversely, overexpose vulnerable groups to harmful or

8 Edelson, Lauinger and McCoy, 'A security analysis of the Facebook Ad Library' (n 7).

9 Thomas Birkner, Erik Koenen and Christian Schwarzenegger, 'A Century of Journalism History as Challenge: Digital archives, sources, and methods' (2018) 6 *Digital Journalism* 1121.

10 Jason Miller, 'Regulating robocalls: Are automated calls the sound of, or a threat to, democracy?' (2009) 16 *Michigan Technology Law Review* 213.

11 Tom Dobber, Ronan Ó Fathaigh and Frederik Zuiderveen Borgesius, 'The regulation of online political micro-targeting in Europe', 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1440> accessed 24 September 2022.

12 Frederik Zuiderveen Borgesius and others, 'Online Political Microtargeting: Promises and Threats for Democracy' (2018) 14 *Utrecht Law Review* 82. William Gorton, 'Manipulating citizens: How political campaigns' use of behavioral social science harms democracy' (2020) 38 *New Political Science* 61.

13 Young Mie Kim and others, 'The stealth media? Groups and targets behind divisive issue campaigns on Facebook'(2018) 35 *Political Communication* 515.

manipulative content.[14] Here, too, personalisation may frustrate the ability to detect and address wrongdoings. Although empirical evidence exists for many of the above claims, a lively debate persists about the overall significance of these microtargeting concerns relative to other policy concerns in media governance.[15]

Seen in this light, the Facebook Ad Library represents a potentially significant shift in the affordances of online microtargeting: by creating a public record of personalised advertising messages, it may help to diagnose and address many the above harms. However, the governance literature on transparency and accountability warns that such assumptions about the salutary effects of information disclosure should be approached critically, and that much depends on whether watchdogs organisations, particularly the media, actually use the available information for accountability purposes.

2.3 The Ad Library as a tool for watchdog journalism

The governance literature emphasises that transparency is not a guarantee for accountability, but merely a precondition.[16] The accountability effects of transparency are not self-executing, but depend on relevant stakeholders to actually use the available information and attach consequences to it.[17] In practice, however, disclosures may lack a 'critical audience' with the capacity and interest to fulfil this role.[18] In the context of online campaigning, Katherine Dommett has therefore warned that 'it is not clear whether citizens are aware of, or could easily discover the existence of, [ad] archives'.[19]

Scholarship routinely asserts the governance benefits of public transparency, but these are almost never tested empirically.[20] What little evidence we do have, mostly

14 Balazs Bodó, Natali Helberger and Claes de Vreese, 'Political micro-targeting: a Manchurian candidate or just a dark horse? Towards the next generation of political micro-targeting research' (2017) 6(4) *Internet Policy Review* <https://policyreview.info/articles/analysis/political-micro-targeting-manchurian-candidate-or-just-dark-horse> accessed 16 September 2022. Dobber, O Fathaigh and Zuiderveen Borgesius (n 11). Muhammad Ali and others, 'Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes' (2019) 3 *Proceedings of the ACM on human-computer interaction* 1.

15 Kim and others, 'The stealth media?' (n 13). Ali and others, 'Discrimination through optimization' (n 14). Yochai Benkler, Robert Faris and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press 2018).

16 Albert Meijer, 'Transparency', in Mark Bovens, Robert Goodin and Thomas Schillemans (eds.), *The Oxford Handbook of Public Accountability* (Oxford University Press 2014).

17 Ibid.

18 Jakko Kemper and Daan Kolkman, 'Transparent to whom? No algorithmic accountability without a critical audience' (2019) 22 *Information, Communication & Society* 2081.

19 Kate Dommett, 'Regulating digital campaigning: The need for precision in calls for transparency' (2020) 12 *Policy & Internet* 432.

20 Igbal Safarov, Albert Meijer and Stephan Grimmelikhuijsen, 'Utilization of open government data: A systematic literature review of types, conditions, effects and users' (2017) 22 *Information Polity* 1.

from the open government context, indicates that most public transparency resources are underused, and almost never consulted by individual citizens.[21] This literature emphasises the importance of mediation by specialised stakeholders who process open data and recirculate its insights to general audiences.[22] Journalists in particular are highlighted as key users of open data and as agents of public accountability.[23] Research by Kate Dommett into the UK media's digital campaigning coverage has already shown that platform disclosure policies, including ad archives, can both enable and constrain reporters on topics of public interest.[24] This paper builds on such findings by focusing the affordances of one specific tool, the Facebook Ad Library, for reporters across different (regional and topical) contexts.

What role do journalists play in public accountability? Formally, journalists have no power to impose sanctions on other stakeholders such as platforms or advertisers. Instead, their reporting can act as a catalyst for other forms of accountability, such as electoral, legal or social accountability.[25] Pippa Norris observes that watchdog journalism can contribute to accountability in two ways: a more concrete primary function of revealing specific instances of malfeasance, and a more diffuse secondary function of informing public deliberation and democratic self-governance.[26] The Ad Library could conceivably contribute to both functions since, as discussed, the opacity of online advertising is associated with both individual wrongdoings and with the barriers to public deliberation. Journalism about the personalised targeting of ads could also constitute what Nicholas Diakopoulous has termed 'algorithmic accountability reporting' which 'seeks to articulate the power structures, biases, and influences that computational artifacts play in society'.[27]

21 Meijer, 'Transparency' (n 16). Alfonso Quarati and Monica de Martino, 'Open government data usage: a brief overview' (2019) *IDEAS '19: Proceedings of the 23rd International Database Applications & Engineering Symposium*.

22 Meijer, 'Transparency' (n 16). Archon Fung, 'Infotopia: Unleashing the Democratic power of transparency' (2013) 41 *Politics & Society* 183. Rui Pedro Lourenço, 'Evidence of an open government data portal impact on the public sphere' (2016) 12 *International Journal of Electronic Government Research* 21.

23 Ibid.

24 Dommett, 'Regulating digital campaigning' (n 19).

25 Pippa Norris, 'Watchdog journalism', in Mark Bovens, Robert Goodin and Thomas Schillemans (eds.), The Oxford Handbook of Public accountability (Oxford University Press 2014). Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 *European Law Journal* 447.

26 Norris, 'Watchdog journalism' (n 25).

27 Nicholas Diakopoulos, 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures' (2014) 3 *Digital Journalism* 398.

To study watchdog journalism empirically, Norris outlines three areas of inquiry: '(1) whether journalists accept their role as watchdogs, (2) whether they act as watchdogs through their coverage in practice, and (3) whether this activity serves as an effective accountability mechanism by mobilising voters, policymakers or other democratic forces'.[28] In other words: attitudes, coverage, and impact. We explore attitudes and coverage through a combination of content analysis and interviews. Taken together, these also provide starting points for the assessment of impact.

3. Content analysis

3.1 Methods

Given the exploratory nature of this research, we first performed an inductive, qualitative pilot study in order to generate a typology of journalistic references to the Ad Library. This provided the basis for a large-scale quantitative analysis of articles via the LexisNexis database. Together, these analyses illustrate the general substance, scale and geographic distribution of Ad Library journalism.

3.2 Pilot study

Our pilot study took place in May 2020. We studied articles referencing the Facebook Ad Library through Google Search and Google News, based on keyword searches for <"Facebook" AND "Ad Library" OR "Ad Archive">. News articles containing concrete references to Ad Library data were selected for analysis. In total we collected 38 such articles. Through qualitative, inductive analysis, we devised a typology of different forms of usage.[29] In particular, our analysis focused on the types of data involved, whether any wrongdoing was asserted, and the norms or standards invoked. This typology was operationalised and refined iteratively into a protocol for large scale quantitative analysis, which we discuss below.

3.3 Content Analysis Protocol

From our pilot study it immediately became clear that many journalistic references to the Ad Library consisted of describing the Ad Library as a phenomenon, rather than actually using the data it offers. Announcements and updates to the Ad Library made headlines regularly as it was updated, expanded, and gradually rolled out across the globe (e.g. 'Facebook Is Taking Steps to Safeguard Canada's Oct. 21 Federal Election'). These articles, which we term 'metacoverage', were filtered out from further analysis since they do not involve any usage of the Ad Library as a tool for transparency ('Non-

28 Norris, 'Watchdog journalism' (n 25).

29 Philipp Mayring, 'Qualitative content analysis' (2000) 1 *Forum: Qualitative Social Research* 159.

metacoverage'). More specifically, we filtered out articles lacking references to actual data from the Ad Library such as concrete spending figures or advertising messages, as well as articles that reference Ad Library data solely to illustrate its affordances.

For articles that actually use the Ad Library, we distinguished between two types following Norris' (2014) aforementioned distinction between the primary and secondary functions of watchdog journalism: calling attention to wrongdoing, and disseminating information in service of public deliberation. We operationalised this distinction by coding whether the article purported to expose any potential wrongdoing related to Facebook advertising cited from the Ad Library, based on criticism supplied by the author or a quoted source ('Wrongdoing reported'). Wrongdoing in this account can include potentially unlawful activity but also anything described as harmful or unethical. Such allegations must be made explicitly in the article by either the author or a quoted source. For instance, reporting on wrongdoing includes articles involving allegations of false or misleading advertisements, voter suppression, foreign interference or violations of campaign finance laws. Ad Library usage without any wrongdoing, our pilot study showed, typically focused on spending trends and messaging strategies for political advertising.

We also coded for three specific subcategories of wrongdoing identified during the pilot study: First, wrongdoing related to advertising content, such as misleading or hateful content ('Wrongdoing category: Content of the advertisement'). Second, wrongdoing related to personalisation practices, such as discriminatory, manipulative or exclusionary targeting ('Wrongdoing category: Personalisation'). Third, wrongdoing related to the identity of the ad buyer and the origin of their funds, such as deceptive or clandestine ad funding schemes (e.g. 'astroturfing'), the involvement of foreign entities, and the violation of election spending restrictions ('Wrongdoing category: Identity of the ad buyer & origin of funds'). As a proxy for the prominence of wrongdoing within the overall article, we code for each category whether the allegation is described solely in the body text or also in the article headline. In addition, we code whether the wrongdoing is described as a potential violation of applicable Laws ('Violation of Law') and Terms of Service ('Violation of Terms of Service'), in order to clarify the norms and sanctions at stake: whether it concerns a more 'soft' form of accountability based on reputation and publicity, or a 'hard' form of accountability grounded in binding norms and sanctions. We also code whether the ad in question is political or non-political ('Political Ad'), which we operationalise as any ad without an apparent commercial purpose, as well as any ad that is described in the article as having been classified as 'political' by Facebook.

Sample

Our sample was collected from the LexisNexis news archive of print media. We performed keyword searches in the publication categories 'Newspapers' and 'Magazines' and for the regions United Kingdom, United States, Germany and the Netherlands. The United Kingdom and the United States were selected due the prevalence of political microtargeting in these countries, as well as the fact that the Ad Library was launched in these countries before any other. Germany and the Netherlands were selected as additional countries with comparable levels of socioeconomic development yet with relatively smaller-scale political microtargeting industries, as well as due to language considerations.

Our sampling used the keywords <'Facebook' AND 'ad library' OR 'ad archive'>. The keywords 'ad library' and 'ad archive' were combined because nomenclature is not consistent across outlets; the New York Times, for instance, tends to use 'archive', and the Washington Post 'Library'. This likely results from Facebook's own inconsistency on the topic: the company initially branded the tool as an 'Archive', but later rebranded to 'Library'.[30] In Germany and the Netherlands we also included the keywords 'Advertentiebibliotheek' and 'Werbebibliothek', respectively, which are local names for the Ad Library. This approach has certain limitations in detecting Ad Library-related journalism that departs from these referencing conventions, which we discuss in Section V. We searched for articles published between May 2018 (when the Ad Library first became operational) and August 2020. This returned a total of 203 articles, excluding 5 duplicates.

Inter-coder reliability

A sample of 58 articles (28% of all articles) was double-coded by two coders to calculate intercoder reliability (Krippendorff's alpha). Table 2 below lists the results. We were not able to calculate Krippendorff's alpha for the variables 'Wrongdoing category: Personalisation' and 'Wrongdoing norm: Violation of Law' since there were not enough cases where these categories applied. A Krippendorff's alpha of .80 is often seen as the norm of strong reliability, and the cut-off point is .67.[31]

30 Grygiel and Sager, 'Unmasking Uncle Sam' (n 7).

31 Daniel Riffe and others, *Analyzing media messages. Using quantitative content analysis in research* (Routledge 2014).

Variable	Krippendorff's alpha
Non-Metacoverage	.87
Political advertisement	1.00
Wrongdoing reported	1.00
Wrongdoing category: Content of the advertisement	.84
Wrongdoing category: Identity of the ad buyer & origin of funds	.86
Wrongdoing norm: Terms of Service violation	.75

Table 2: Inter-coder reliability scores

3.4 Findings

Our results show that the Facebook Ad Library was referenced in at least 203 print newspaper and magazine articles in the selected countries. The bulk was published in the United States and the United Kingdom, with 150 and 29 articles respectively, compared to Germany's 15 and the Netherlands' 9 (see Figure 2). The list of publishers is likewise dominated by U.S. outlets, with only one U.K. outlet breaking the top 5: AdWeek (33 articles), the New York Times (31 articles), CE Notificias Financeras (25 articles), the Washington Post (21 articles), and The Guardian (17 articles). Together, they account for 62% of our findings, with the remainder being supplied by 48 other outlets. It is worth noting that many of the non-U.S. publications in our sample were in fact reporting about U.S. advertisements, particularly in the United Kingdom and particularly for stories that actually identified potential wrongdoing (discussed below).

In terms of substance, 118 out of 203 articles in our sample, or 58%, consist of metacoverage that merely describes the tool rather than actually using the data on offer (see Figure 1).

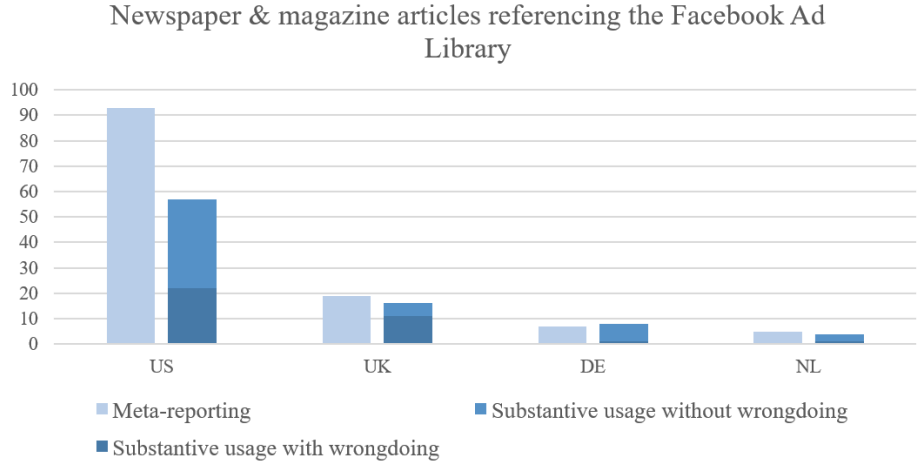

Figure 1: Newspaper & magazine articles referencing the Facebook Ad Library

As for the articles that do use Ad Library data, 50 out of 85 do not allege any particular wrongdoing based on this data (see Figure 1). As discussed, these articles typically focus on campaign coverage, for instance reporting on aggregate spending trends (e.g. 'Biden Pours Millions Into Facebook Ads, Blowing Past Trump's Record') or messaging (e.g. 'Trump Campaign Facebook Ad Strategy: Paint Biden As A Socialist').[32] Although the majority appears to focus on election and referendum campaigns, other issues are also reported on occasionally. To take one notable example, the Washington Post cited the Ad Library to report on the FBI's use of Facebook ads to recruit Russian informants.[33] Ad Library usage is not always central to the article's topic, but can also be used more incidentally as context for other stories.

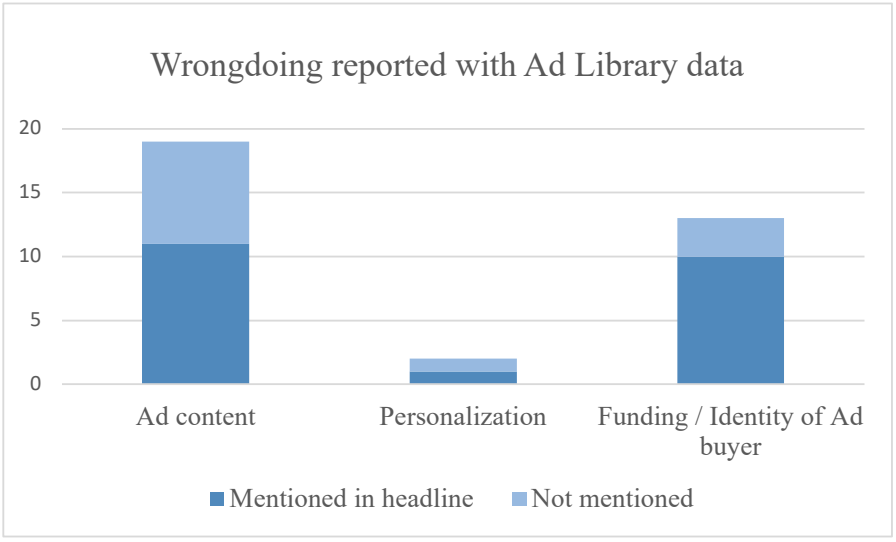

Figure 2: Wrongdoing reported with Ad Library data

As for articles about possible wrongdoing, 19 counts related to the content of the ad, 13 to the identity of the ad buyer, and only 2 to personalisation. Here it bears repeating that the Ad Library provides only limited information on personalisation techniques.

32 Shane Goldmacher, 'Biden Pours Millions Into Facebook Ads, Blowing Past Trump's Record', *The New York Times* (2 September 2020) < https://www.nytimes.com/2020/06/08/us/politics/biden-trump-facebook-ads.html> accessed 26 September 2022. Jack Brewster, 'Trump Campaign Facebook Ad Strategy: Paint Biden As A Socialist', *Forbes* (13 April 2020) < https://www.forbes.com/sites/jackbrewster/2020/04/13/trump-campaign-facebook-ad-strategy-paint-biden-as-a-socialist/?sh> accessed 25 September 2022.

33 David Cohen, 'FBI Uses Facebook Ads in Washington, D.C., to Find Information on Russian Spies' *Adweek* (2 October 2019) <https://www.adweek.com/performance-marketing/fbi-uses-facebook-ads-in-washington-d-c-to-find-information-on-russian-spies/> accessed 26 September 2022.

Just under two thirds of these articles (22/34) mention these issues in the headline as well as in the body text. A possible violation of the law was alleged in 4 cases. In 9 cases, the ads were also described as violating Facebook's Terms of Service.

Content-based wrongdoings mostly involved allegedly false or misleading statements. Accusations of hate speech were also found, for instance regarding Trump advertisements involving alleged Nazi symbols.[34] Wrongdoings related to the identity of the ad buyer typically focused on the misleading use of astroturf groups and the listing fake or misleading names in relevant disclosures to Facebook (e.g. 'In Virginia House Race, Anonymous Attack Ads Pop Up on Facebook').[35]

Only two articles in our sample involved commercial ads: one about false ads for solar panels and another about misleading ads for HIV medicine. Here it bears repeating that the Ad Library's functionalities for commercial ads are restricted substantially compared to political ads.

4. Interviews

Building on the above content analysis, we also interviewed journalists to discuss their experiences in using the Ad Library. Recalling Norris' three empirical aspects of watchdog journalism—attitudes, coverage, and impact—the above content analysis demonstrates coverage, and these interviews allow us to explore attitudes.[36] Extensive survey research has already examined the self-conception of journalists as public watchdogs in a general sense (e.g. Weaver et al 2007), but no research has yet focused on their perception of the Ad Library as a means to this end. Our aim here is not to develop claims that generalise across journalism writ large, but rather, through qualitative, in-depth interviews to unpack the particular motives and experiences of journalists who have used the Ad Library.[37]

4.1 Method

We approached 16 journalists with experience using the Ad Library, and 12 of them agreed to be interviewed. Participants were selected based on published work

34 Scott Nover, 'Facebook Removes Trump Campaign Ads for Including Symbol Used by Nazis' *Adweek* (18 June 2020) < https://www.adweek.com/programmatic/facebook-removes-trump-campaign-ads-symbol-nazis/> accessed 26 September 2022.

35 Kevin Roose, 'In Virginia House Race, Anonymous Attack Ads Pop Up on Facebook', The New York Times (17 October 2018) < https://www.nytimes.com/2018/10/17/us/politics/virginia-race-comstock-wexton-facebook-attack-ads.html> accessed 26 September 2022.

36 Norris, 'Watchdog journalism' (n 25).

37 Grant McCracken, *The Long Interview* (Sage Publications 1988).

identified in the content analysis pilot study, combined with snowball sampling. Our aim in sampling was to obtain a diversity of perspectives, in terms of participants' location, venue, and beat. We prioritised journalists with multiple publications based on the Ad Library, but also included several with only one or two relevant publications. Due to language considerations, our selection was limited to journalists working in English or Dutch. Interviews were conducted in the period September-November 2020 via Zoom videoconferences. Table 2 provides an overview of participants and their titles and affiliations at the time of our interviews, which we publish with their permission. In some cases, relevant work was published on a freelance basis, or with a former employer; these outlets are listed in brackets.

Name	Title	Outlet
Coen van de Ven	Investigative Journalist	De Groene Amsterdammer
Mark Scott	Chief Technology Correspondent	Politico
Madelyn Webb	Investigative Researcher	First Draft
Ryan Mac	Senior Technology Reporter	The New York Times (published in Buzzfeed News)
Nick Garber	Reporter	Patch (published in Pennsylvania Post-Gazette)
Eric van den Berg	Investigative Journalist	Freelance (published in Brandpunt)
Josh Keefe	Investigative Reporter	Bangor Daily News
Reinier Kist	Media Editor (*Redacteur*)	NRC
Jeremy B. Merrill	Investigative Data Reporter	The Washington Post (also published in ProPublica, the Markup)
Matt Novak	Senior writer	Gizmodo
Rik Wassens	Data Journalist and Editor (*Redacteur*)	NRC
Kayla Gogarty	Senior Researcher	MediaMatters

Table 3: Overview of interviewees

The interviews were conducted in a semi-structured format, with an interview guide based on the following questions:

1. Use cases: How, if at all, has the Ad Library appeared in your work?
2. Research processes: Could you describe your process in using the Ad Library?
3. Attitudes: What is your opinion on the usefulness of the Ad Library as a tool for journalists?
4. Outlook: Do you intend to use the Ad Library in future?

We discuss our findings in the corresponding order.

4.2 Findings

Use cases

The use cases mentioned by participants mirrored the results of our content analysis: participants were able to make good use of spending and content data, but lamented the lack of targeting data. In addition, participants also highlighted the role of the Ad Library as a means to evaluate the enforcement of Facebook's own policies, such as ad pricing and content rules.

Reporting on ad spending and funding sources: A majority of the journalists we spoke to (7/12) highlighted the Ad Library's insights into ad spending, for instance as a means to 'follow the money' (Coen van de Ven) or 'to see who's been spending what and how' (Mark Scott). Madelyn Webb, a disinformation researcher, used the Ad Library because 'there's interesting stories to be told about who is spending money on particular misleading narratives.' Rik Wassens recounted that his editors also emphasised the significance of spending in their headlines: 'The amount of money. That is absolutely the most newsworthy. I don't write the headlines myself but they do show you how the institutions view things, and there you go: 'Socialist Party sends €50,000.'

Participants also highlighted the role of the Ad Library in detecting new actors and sources of funding. According to Jeremy B. Merrill, '[w]hat's interesting about these ads from the Ad Library–it's not often *that* the ad ran, it's who this group is that now exists. It's some group that you've never heard of before, that's running ads. […] The story then is that there's this group, and that they're spending 10,000 dollars or whatever.' Examples from participants include anti-union astroturfing groups, lobbying funded by energy and fossil fuel companies, as well as propaganda from Chinese and Turkish state media related to the oppression of Uighurs and Kurds.

A specific use case for spending data highlighted by Jeremy B. Merrill was researching Facebook's differential pricing policies: by comparing spend and view data per campaign, he was able to show that the Biden campaign had been charged higher rates on average than the Trump campaign. Mark Scott described how the Ad Library helped to uncover unlawful campaign finance practices in the United Kingdom: 'The dollar or euro spend in specific races has been interesting because it provides a clear example, for instance in the U.K. 2019 election, of people breaking the law: candidates using money in their constituency that they weren't supposed to.'

Journalists nonetheless faced important obstacles in researching ad spending through the Ad Library. The data was insufficiently granular, since it is disclosed in general ranges rather than precise amounts. Participants also reported difficulties

in overseeing spending by entities with multiple Facebook pages and accounts. Furthermore, the names provided under 'paid for' disclosures were often imprecise, referring to non-existent organisations or proxy organisations. In these instances, the Ad Library merely served a starting point for investigation, and other forms of research were necessary to uncover, if possible, the true origin of funds. In the words of Nick Garber, 'I had to do some digging to understand that the group that was named as the sponsor had ties to a much larger parent organisation.' Likewise, Jeremy B. Merrill recounted: 'I had to do a whole bunch of shoe-leather reporting to figure it out'.

Reporting on targeting practices: Almost all participants expressed interest in targeting practices, and complained that the Ad Library failed to offer meaningful information about this issue (8/12). The Ad Library offers highly generic reach data and no concrete information as to the targeting mechanisms involved. Only in exceptional cases could this reach data be used to infer targeting strategies, according to Coen van de Ven:

> Where does it deviate? With all the political party data you tend to see a certain distribution in terms of age, location, and it's almost never surprising. And so if there's an exception, that's when I start paying attention. That's when I think: How can that be? If I see a 100% female reach, then I know: this wasn't targeted at men. That's an assumption I'm allowed to make. […] So I'm happy that it exists, it's better than nothing. But I'm still missing a lot. It's not the transparency we as journalists or other researchers were hoping for.

Researchers also noted that targeting strategies could sometimes be inferred from advertising content. For instance, an advertisement about 'Latinos4Trump' can be assumed to be aimed at a certain demographic, although the precise targeting mechanisms remain uncertain. Barring such exceptions, however, the lack of targeting data surfaced as one of most frequently and strongly voiced criticisms of the Ad Library, and as one of the acute constraints on the types of reporting that this tool allows journalists to pursue.

Reporting on ad contents and content policy enforcement: Other use cases related to the content of advertisements, and, relatedly, how Facebook enforces their content policies. Media watchdog researchers used the Ad Library regularly to search for harmful content. Kayla Gogarty from Media Mattters described her routine as follows: 'I basically have sets of pages that I would follow almost on a daily basis, particularly to look for repeat offenders—accounts that we know will frequently post misinformation in their ads.' Two participants used to the Ad Library to detect manipulated media, by cross-referencing ad content with original sources. Gogarty also recounted assisting

other colleagues at Media Matters in using the Ad Library, for instance helping their LGBTQ programme team to trace the spread of anti-trans Facebook advertising.

Related to the above, several journalists (4/12) described how they used the Ad Library to detect gaps and inconsistencies in Facebook's Terms enforcement. As Madelyn Web of First Draft put it: 'Every time they say they're gonna take something down, we can find examples of it. […] When they announced the QAnon takedown, I was like 'hmm, okay', so I went to the Ad Library'. Kayla Gogarty: 'If there's a new Facebook policy that's coming out, I'll go and check: are these ads not following this policy, might they have slipped through the cracks?' Ryan Mac considered holding Facebook accountable his primary use case for the Ad Library: 'What I do is corporate accountability. It's not necessarily holding up a press release about what the company's doing, it's: here's what the company says it's going to do, and here's what it's doing wrong.' For instance, he used the Ad Library to show that Facebook had enforced its rules on clickbait inconsistently, allowing Trump to run numerous ads that violated the company's policies. 'It's the Facebook policy, and so you want it to be applied equally across something as consequential as the U.S. election. If it's not, that's giving a candidate by definition an unfair advantage. And that's a story.'

Research processes

We discussed how journalists discover relevant information in the Ad Library. Participants engaged in both proactive research consisting of browsing or analysing Ad Library data, as well as reactive research prompted by third-party tips. One telling example comes from Nick Garber, who was directed to the Ad Library by a labour union representative he interviewed, ultimately leading him to discover an anti-union influence network. New policy announcements from Facebook were another important prompt to check the Ad Library. Ryan Mac, one of the most prolific reporters of Ad Library stories in our sample, put it as follows: 'I'm sure there's a reporter out there who checks the Ad Library every day, but for most journalists it's a sporadic thing that they'll check now and then when they have a tip.' Others made a habit of searching the Ad Library more regularly. Indeed, two of the journalists we spoke to, Madelyn Webb and Matt Novak, did claim to check the Ad Library every morning, at least during elections.

Most participants lacked the expertise to make use of the API themselves (10/12), and either stuck to the browser interface (6/12) or enlisted the help of specialists to gather data through the API (4/12). Two journalists in our sample preferred to work with data collected independently through volunteers with browser plugins, which automatically collect data about the advertisements shown to individual participants as they browse the web, rather than relying solely on Ad Library data. They used the Ad Library mostly

to enrich and corroborate their independent observations. For instance, Jeremy Merrill described his involvement in the NYU Ad Observatory, which combines data from browser plugins and the Ad Library with a view to supporting journalists in reporting on online political advertising.

Attitudes towards the Ad Library

We discussed how participants perceived the usefulness of the Ad Library as a tool for journalists. The responses indicated a love/hate relationship: the Ad Library was considered an improvement over the default opacity of political microtargeting, but most participants remained sharply critical of its flaws and shortcomings. Only two participants had no particular criticisms of the tool, and these were both once-off users who did not use the Ad Library regularly.

The perceived advantages of the Ad Library related to the use cases it enabled, described above, such as corporate accountability and the combating of disinformation. Two participants articulated a more general desire to bring visibility to personalised messaging, and exposing it to public deliberation and scrutiny. Eric van den Berg remarked:

> What always surprises me is that a lot of things happen which reach a very large audience, but still seem to enjoy a kind of relative invisibility. Journalists don't write about it, and as a result the standards for what constitutes normal behaviour seem to be very different. I think the things that the VVD [The Netherlands' incumbent political party] gets up to on Facebook would lead to shocked reactions in Parliament.

Similarly, Madelyn Webb recounted that 'it feels a little backdoor, a little salacious, so journalists like it. […] It feels a bit like being a private investigator, or like doing FOIAs. It feels like a scoop, even though a lot of people are seeing it.'

Against these benefits, participants offered many criticisms of the Ad Library. Most common was the lack of targeting information, discussed previously. The lack of granularity in both reach and spending was also a recurring theme. Mark Scott, who reported on elections in both the United States and Europe, highlighted that European versions of the Ad Library were even less detailed than the U.S. version. Participants also criticised the reliability and user-friendliness of both the API and the browser tool. Tracking overall campaign spending was difficult since platforms presented spending data per Page, whereas campaigns often operated multiple Pages. Two participants also expressed concerns that the data risked misleading non-expert journalists, who might for instance mistake reach data for targeting data, or Page spending for total campaign

spending. Another drawback journalists mentioned was that Ad Library research was time-consuming, and difficult to accommodate in their busy schedules.

Outlook

Most participants were interested in continuing to use the Ad Library, though few had concrete ideas or plans. Participants from the Netherlands already intended to continue using the tools for the upcoming elections of Spring 2021, and predicted that attention for the tool would increase as Facebook advertising grew in scale and significance for domestic political campaigns. Participants also described how they were helping to make the Ad Library more visible and accessible amongst their peers, for instance by organising public webinars for investigative journalists, internal seminars for newspaper colleagues, Twitter bots repurposing Ad Library data, and the aforementioned NYU Ad Observatory.

5. Discussion

Our paper confirms that the Facebook Ad Library has supported watchdog journalism. We find evidence for both the primary watchdog function of calling attention to wrongdoing by powerful actors (in this case: Facebook and its advertisers) as well as the secondary watchdog function of disseminating general information about public affairs (in this case: microtargeted political campaigns). As regards the primary watchdog function, these stories tend to revolve around calling attention to influence networks and astroturf operations, as well as monitoring ad content for hate speech and disinformation. Another recurring issue was the consistency and fairness of platform policy enforcement, especially their content policies but also other aspects such differential pricing between campaigns. Turning to the secondary watchdog function, these stories tended to focus on campaign reporting, in particular on spending trends and to a lesser extent messaging and targeting strategies. In addition to these recurring themes, the Ad Library also featured in a range of more unexpected and niche topics, from FBI recruitment ads to scams targeting elderly Trump voters.

We observe a notable geographic discrepancy in Ad Library usage: it is most prevalent in the US, less so in the United Kingdom and less still in Germany and the Netherlands. It goes beyond the scope of this paper to offer an exhaustive explanation for these discrepancies, but the most readily apparent factor seems to be that political microtargeting simply takes place on a far larger scale in the United States.[38] In the

38 Dobber, Ó Fathaigh and Zuiderveen Borgesius, 'The regulation of online political micro-targeting in Europe' (n 11).

Netherlands and Germany, by contrast, political advertising budgets are only a fraction of those in the US, and it stands to reason that the issue does not receive the same level of attention. Of course, these circumstances may change. Interview participants from the Netherlands predicted that online advertising would increase in future elections, as would usage of the Ad Library.

We found more metacoverage about the Ad Library than actual usage. Arguably, this suggests that this tool has been successful for Facebook at least as a PR measure, generating coverage about their efforts to create transparency, without necessarily receiving scrutiny of the practices at issue. Still, we do find evidence that such scrutiny takes place at least in some cases, leaving it up for discussion whether the public interest value of this watchdog journalism justifies the accolades that we find Facebook to have received.

As for attitudes, our interviewees' opinions on the Ad Library might be summarised as 'better than nothing'. They perceived a strong public interest in public advertising transparency, and considered the Ad Library a significant improvement over the *status quo ante* of total opacity. However, most participants remained sharply critical of the numerous shortcomings in the Ad Library's present implementation, including but certainly not limited to the lack of targeting data and the lack of user-friendliness. Many of the most specialised journalists still preferred to work with alternative data collection methods such as data scraping via browser plugins. But Facebook has recently started cracking down on these independent collection methods, leaving journalists all the more reliant on the inferior offerings of the Ad Library.[39]

5.1 Impact: From publicity to accountability?

Publicity does not guarantee accountability. The power of the press over platforms and their advertisers is indirect and contingent on its (perceived) ability to mobilise an effective response from other stakeholders, such as end users, voters, governments or regulators. Given that dominant platforms such as Facebook are able to act with relative impunity towards many of these stakeholders, watchdog journalism might too be 'disconnected from power'—as transparency measures often are.[40] How, and when, might it make a difference?

39 Andrew Sellars, 'Facebook's threat to the NYU Ad Observatory is an attack on ethical research', *NiemanLab* (29 October 2020) <https://www.niemanlab.org/2020/10/facebooks-threat-to-the-nyu-ad-observatory-is-an-attack-on-ethical-research/> accessed 19 September 2022.

40 Patrick Barwise and Leo Watkins, 'The evolution of digital dominance: how and why we got to GAFA', in: Martin Moore and Damian Tambini (eds.), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (Oxford University Press 2018).

The strongest evidence of accountability we find in cases where the advertising is alleged to violate formal rules, such as platform Terms of Service and/or applicable laws. These cases typically lead to removal of the ads in question, and could also trigger legal action. The causal chain from information disclosure to repercussion is relatively short and accountability thus relatively plausible and tangible.

In other cases, our findings are at most suggestive of softer, more diffuse forms of political accountability. Straightforward campaign reporting that does not involve any particular wrongdoing could still conceivably contribute to electoral accountability of campaigners towards voters, by exposing targeted campaigns to a broader public and thus mitigating the 'fragmentation' of political campaigning associated with political microtargeting.[41] This is especially likely in cases where reporters aim to highlight inconsistencies in messaging towards different constituencies, such as Matt Novak's article at GizModo highlighting that 'Trump's New Facebook Ads Claim He's Peacenik Who Also Loves Assassinations'.[42] Other articles do not address consistency as explicitly, but could still have some plausible constraining or 'defragmenting' effects. For instance, reports observing that 'Trump's deluge of Facebook ads have a curious absence: coronavirus' can be conceived of as catalysing a more informed public discourse about the priorities of this campaign, a form of public accountability which might feed into any number of more proximate accountability processes.[43] In addition to electoral accountability towards voters, this campaign reporting could conceivably catalyse other forms of social and political accountability, for instance by spurring legislative or regulatory reforms.[44] As an empirical matter, more detailed process tracing would be needed to demonstrate any such effects conclusively.

It is worth noting that both disinformation and astroturfing, two of the most common forms of wrongdoing identified through the Ad Library, are not always prohibited by Facebook or by the law. Here too, watchdog journalism depend on 'soft' forms of public accountability. In principle, journalistic fact-checking of microtargeted ads could offer a direct corrective to disinformation in the minds of citizens, but a growing

41 Zuiderveen Borgesius and others, 'Online Political Microtargeting: Promises and Threats for Democracy' (n 12).

42 Matt Novak, 'Trump's New Facebook Ads Claim He's Peacenik Who Also Loves Assassinations', *Gizmodo* (7 August 2020) <https://gizmodo.com/trumps-new-facebook-ads-claim-hes-peacenik-who-also-lov-1844644446> accessed 27 September 2022.

43 Julia Carrie Wong, 'Trump's deluge of Facebook ads have a curious absence: coronavirus', The Guardian (26 March 2020) < https://www.theguardian.com/us-news/2020/mar/26/trump-facebook-ads-immigrants-coronavirus> accessed 27 September 2020.

44 Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 *European Law Journal* 447.

empirical literature raises questions about the efficacy of this approach.[45] If reporting on such issues is to have any accountability effect, therefore, it depends primarily on its capacity to catalyse a response from governments, platforms or other influential actors in advertising governance.

The journalism we describe here does not fit neatly in the bucket of 'algorithmic accountability reporting'.[46] Our content analysis shows that algorithmic personalisation rarely features in Ad Library journalism, and instead points towards other aspects of platform advertising besides algorithmic decision making that warrant transparency in their own right, such as ad content, spending and buyer identities.[47] Our interviews clarify that this lack of algorithmic accountability reporting is certainly not for a lack of journalistic interest—many of our participants were in fact eager to investigate targeting practices—but rather a lack of data access. As we discuss further below, this illustrates clearly how Facebook's disclosure policies constrain and shape reporting practices.

5.2 Critical perspectives on Ad Library journalism

Having discussed some of the Ad Library's benefits, we now turn to more critical reflections. Firstly, the public interest value of Ad Library journalism is not given but debatable. Two of the journalists we spoke to already raised tentative questions about merely descriptive campaign reporting based on Ad Library spending data; was this not so much more 'low-hanging fruit' or 'horse-race coverage'?[48] Particularly where Ad Library reporting merely restates aggregate spending data without further contextualisation or analysis, the public interest value of this reporting need not be overstated. Indeed, besides the high-minded ideals of watchdog journalism, more mundane considerations such as mere novelty and availability may also factor into Ad Library usage.

Secondly, the Ad Library's limitations and inaccuracies may even pose risks to journalism. First, its data may not always be reliable; for instance, journalists depend on Facebook to identify political ads even though we know this process to

45 Nathan Walter and others, 'Fact-Checking: A meta-analysis of what works and for whom' (2020) 37 *Political Communication* 350.

46 Diakopoulos, 'Algorithmic Accountability' (n 27).

47 Chapters 2 and 3 above (Paddy Leerssen, 'The soap box as a black box: regulating transparency in social media recommender systems' 11(2) (2020) European Journal of Law and Technology <https://ejlt.org/index.php/ejlt/article/view/786>; Paddy Leerssen, Jef Ausloos, Brahim Zarouali, Natali Helberger and Claes de Vreese, 'Platform ad archives: promises and pitfalls' (2019) 8(4) Internet Policy Review <https://doi.org/10.14763/2019.4.1421>).

48 Toril Aalberg, Jesper Strömbäck and Claes de Vreese, 'The framing of politics as strategy and game: A review of concepts, operationalizations and key findings' (2012) 13 *Journalism* 162.

be inaccurate. Second, the available data may divert attention away from other, unauthorised topics, such as reporting on microtargeting or on non-advertising content. Recalling Facebook's ongoing crackdown on independent data collection, they appear to be pursuing a carrot-and-stick approach in which, through the selective granting and withholding of relevant data, reporting is confined to approved topics. In this light, our research illustrates and underscores the concern, also voiced by others including Dommett, that their control over public transparency resources may help platforms to exercise undue influence on journalistic agendas.[49]

A related concern, voiced by several participants, is that the presentation and ordering of the Ad Library's data could mislead reporters, especially non-experts. Given that the journalists we spoke to tended to be highly critical of the Ad Library and aware of its shortcomings, the risk of deception may not be particularly acute at present. It could exacerbate in future, should the Ad Library become more popular amongst a broader set of journalists. Academic partnerships may have a role to play here: projects such as the NYU Ad Observatory and the University of Amsterdam *Verkiezingsobservatorium* now seek to assist non-expert journalists in using the Ad Library.

5.3 Limitations

It bears repeating that our sample of LexisNexis articles does not capture all journalistic usage of the Ad Library, and therefore understates the overall scale of this phenomenon. The most fundamental limitation of our approach is that our sample does not include online journalism, which is appreciable but more difficult to operationalise in any consistent or comprehensive fashion. Indeed, our interviews and pilot study indicate that certain online, tech-focused outlets are particularly frequent users of the Ad Library, including ProPublica, The Markup, and Buzzfeed News. Even broader conceptions of journalism might also consider Ad Library usage by NGOs and activist groups, such as the widely-cited research by U.K. think-tank InfluenceMap about oil and gas companies advertising on Facebook.[50] Our analysis of print media, then, is by no means exhaustive of the journalism in this space, but should merely be seen as indicative of its general order of magnitude, geographical distribution, and composition.

A related limitation is that our keyword-based sampling does not capture usage which does not reference the Ad Library explicitly. We did not cover reporting that neglects

49 Katharine Dommett, 'The inter-institutional impact of digital platform companies on democracy: A case study of the UK media's digital campaigning coverage' (2021) *New Media & Society* <https://doi.org/10.1177/14614448211028546> accessed 25 September 2022.

50 InfluenceMap, 'Big Oil's Real Agenda on Climate Change' (Research Report InfluenceMap 2019). <https://influencemap.org/report/How-Big-Oil-Continues-to-Oppose-the-Paris-Agreement-38212275958aa21196dae3b76220bddc> accessed 19 September 2022.

to cite the Ad Library, uses non-standard nomenclature such as 'public database', or simply cites 'Facebook' as a generic source. One might expect this approach to bias our content analysis towards metacoverage, on the theory that metacoverage is more likely to explicitly refer to the Ad Library by name. However, with supplemental testing we detected no such bias. As detailed in Appendix II, alternative keywords such as <"facebook" AND "political ads"> still return comparable rates of metacoverage.

Related to the above, there may be instances where the Ad Library surfaced an initial lead for journalists, even if it did not feature as a source in any ultimate publication. For instance, Washington Post reporter Nitasha Tiku recounted on Twitter how she started reporting on Facebook's pharmaceutical advertising policies after she 'fell into a Facebook Ad Library rabbit hole'.[51] The Ad Library is not used as a source in the published article, but it did start Tiku towards a newsworthy investigation.

Finally, we have not yet charted in detail the interaction between journalists and other researchers in this space. Numerous stories in our sample did not rely on original journalistic research, but instead originated from academic studies of the Ad Library. Accordingly, our sample may somewhat overstate the degree to which journalists actually use the Ad Library, rather than reporting on Ad Library research conducted by others such as academics.

6. Conclusion

This article has shown that, for all its flaws, the Ad Library has started to find uptake in journalistic practice. Our findings may serve as both an encouragement and a warning.

On the one hand, we have shown how the Ad Library has enabled new forms of watchdog journalism about online ad campaigns and, in some instances, wrongdoings such as hate speech, disinformation, and astroturfing. Even where no particular wrongdoing is uncovered, this reporting could conceivably strengthen public deliberation in and about microtargeting practices. These findings lend empirical

51 Quote from: Twitter.com <https://twitter.com/nitashatiku/status/1234891011555385347> accessed 26 September 2022 ("I started looking into this after a source got female Viagra ads on an Instagram account where she posts no content/has no photo. Then I fell into a Facebook Ad Library rabbit hole: ads for ADHD (Vyvanse!), HIV, cancer, depression, weight loss."). The article in reference: Nitasha Tiku, 'Facebook has a prescription: More pharmaceutical ads', *The Washington Post* (4 March 2020) <https://www.washingtonpost.com/technology/2020/03/03/facebook-pharma-ads/> accessed 26 September 2022.

weight to the rationale of public ad archives as a tool for public accountability, and underscore the role of journalists in realising these goals.

On the other hand, the growing reliance on this tool by journalists also poses risks. First, the data shared by Facebook has been shown to be incomplete and inaccurate, and could potentially mislead journalists. Second, this new resource may also divert attention from issues that Facebook refuses to document in similar detail, such as their targeting practices and non-advertising content. Indeed, given that most articles did not report on any particular wrongdoing, but instead consisted of either relatively uncritical campaign reporting or, even more commonly, coverage about the Ad Library itself, it could be argued that this tool has received outsized attention relative to the actual watchdog journalism it has enabled.

This research has several implications for the regulation of ad archives, as is now being prepared in various jurisdictions. Given that journalists are starting to rely on this data, ensuring its accuracy, comprehensiveness and consistency is all the more urgent. At the same time, our findings underline that this issue may be less critical in countries where political microtargeting is less prevalent compared to hotspots such as the United States and the United Kingdom.

Future research might build on these findings in various ways. As mentioned, more detailed process tracing could help to demonstrate how and when reporting on online ads triggers accountability effects in particular instances. Usage by other groups besides journalists also merits attention, such as by rival campaigners, consumers, commercial entities, regulators and courts. From teenagers trawling the Ad Library for discount codes,[52] to courts and parliamentary committees citing it as evidence,[53] our newfound public access to personalised advertising campaigns may have wide-ranging consequences, which this article has only begun to chart. More generally, future research might examine other tools through which platforms structure access to their data, such as CrowdTangle, and how these affect our capacity for public accountability.

52 Andrew Griffin, 'Tiktok user reveals ingenious Facebook trick to find hidden discount codes', The Independent (2020, August 11) <https://www.independent.co.uk/life-style/gadgets-and-tech/news/discount-codes-facebook-ad-library-tik-tok-a9665221.html> accessed 19 September 2022.

53 *Campaign Legal Center v. Federal Elections Commission* (2020) Case 1:20-cv-00588 (Complaint for declaratory and injunctive relief). Grygiel and Sager, 'Unmasking Uncle Sam' (n 27).

CHAPTER 5

An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation

Abstract

This paper offers a legal perspective on the phenomenon of shadow banning: content moderation sanctions which are undetectable to those affected. Drawing on recent social science research, it connects current concerns about shadow banning to novel visibility management techniques in content moderation, such as delisting and demotion. Conventional moderation techniques such as removal or account suspension can be observed by those affected, but these new visibility restrictions often cannot. This lends newfound significance to the legal question of transparency rights for content moderation, which are now being proposed in the EU Digital Services Act (DSA). Its new due process framework for content moderation, I show, prohibits shadow banning with only limited exceptions. Through these exceptions, the DSA aims to delineate two competing models for content moderation: as rule-bound administration or as adversarial security conflict.

The more fundamental challenge for this regime will be to define the boundaries of content moderation itself, and to distinguish it from more systemic modes of content curation. Responding to claims that demotion is entirely relative, and therefore not actionable as a category of content moderation sanctions, I show how visibility restrictions can still be regulated when defined as ex post adjustments to engagement-based relevance scores. Understood in this way, safeguards against demotion may help to regulate shadow banning, and shed light on relatively fine-grained and targeted exercises of platform ranking power. Still, these safeguards will not be exhaustive of ranking power, as it is exercised not only through individual cases of moderation, but through structural acts of content curation; not just by reducing visibility, but by producing visibility.

1. Introduction

Content moderation knows no shortage of scandals. From Twitter suspending Donald Trump to YouTube banning Alex Jones, the public record is rife with controversy, debate, and backlash. And yet, speculation abounds that many more cases may be hidden from view. 'Shadow banning', as it has come to be known, alleges that platforms intervene in subtler ways, not by suspending users outright but by secretly demoting them in their recommender systems.

Accusations of shadow banning trigger conflicting responses. For some, it is mere paranoia, stemming from misunderstandings about the ways platforms curate content. Others agree that shadow banning exists, but disagree as to its merits. Is it devious and undemocratic subterfuge, repugnant to fundamental rights and due process? A harsh but necessary defence against social media's most persistent bad actors? Or simply an unintended by-product of new visibility management techniques in content moderation? These questions have become all the more pressing as the EU moves to regulate due process and transparency for content moderation in its new Digital Services Act. This law attempts to settle the shadow banning question: when, if at all, should content moderation decisions be allowed to remain secret?

This paper offers a legal perspective on the shadow banning phenomenon. Drawing on recent social science research, it discusses shadow banning in terms of its terminology, techniques, and policy drivers. Then it examines how the DSA regulates shadow banning through its due process framework for content moderation, and how it attempts to balance conflicting interests in transparency and secrecy. The final section critiques the concept of 'demotion', which is central to both the shadow banning imaginary and the DSA's response to it. I review the challenges in defining and enforcing demotion as a category of content moderation actions, and its limitations in checking the more structural dimensions of recommender governance as a a means of content curation.

2. 'Shadow banning' as a function of visibility remedies

This section introduces the shadow banning phenomenon. It discusses the concept of shadow banning, the content moderation techniques involved, and the policy considerations behind it. My core claim is that shadow banning refers primarily to output-based forms of opacity: the *what* of moderation, not the *why*. In terms of outputs, conventional takedown methods are self-evident to those affected, and therefore do not

afford effective shadow banning. Visibility remedies, by contrast, are output-opaque by default, and act as shadow bans unless expressly notified. For this reason, the growing reliance on visibility remedies threatens to make content moderation more opaque, and lends renewed urgency to the regulation of notification safeguards.

2.1 Definitions: what is 'shadow banning'?

The term 'shadow banning' is colloquial in origin and its usage has changed over time. Originally, the term referred to a deceptive type of account suspension on web forums: a shadow banned user would be give the impression that they were still able to post, whereas in fact their content was no longer visible to any other users.[1] Some sources continue to use the term in this way (including, as we will see, the DSA). But in more recent usage, shadow banning usually refers to alternative remedies, especially visibility remedies such as delisting and downranking.[2] These remedies do not cut off access to content entirely, but instead make it less visible through content discovery features such as search and recommendation.

It is in this new, broader form, that shadow banning has become prevalent in popular and academic discourses. In 2018, US president Donald Trump accused social media firms (without evidence) of 'shadow banning' conservative viewpoints.[3] Elon Musk, during his attempt to takeover of Twitter, tweeted ominous imagery of a shadowy cabal he described as the 'Twitter Shadow Ban Council'.[4] On the other end of the political spectrum, shadow banning allegations have also been raised by marginalised groups including online sex workers and LGBT+ users, as well as by climate activists.[5]

1 Courtney Radsch, 'Shadowban / Shadow-ban', in: Luca Belli, Nicolo Zingales and Yasmin Curzi (eds). *IGF Glossary of Platform Law and Policy Terms* (Internet Governance Forum 2022) <https://platformglossary.info/> accessed 19 September 2022.

2 e.g. Kelley Cotter, '"Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms' (2021) *Information, Communication & Society* <https://doi.org/10.1080/1369118X.2021.1994624> accessed 15 September 2022.

3 Radsch, 'Shadowban / Shadow-ban' (n 1).

4 Gabriel Nicholas, 'Shadowbanning Is Big Tech's Big Problem', *The Atlantic* (28 April 2022) <https://www.theatlantic.com/technology/archive/2022/04/social-media-shadowbans-tiktok-twitter/629702/> accessed 19 September 2022.

5 Carolina Are, 'The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram' (2021) 0 *Feminist Media Studies* 1. Rachel Griffin, 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality', SSRN Draft Paper (2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064738> accessed 19 September 2022. Josephine Lulamae, 'How the "Shadow Banning" Mystery is Messing with Climate Activists' Heads', *AlgorithmWatch* (11 February 2022) <https://perma.cc/JYC5-BQ82> accessed 28 June 2022. Cotter, 'Shadowbanning is not a thing' (n 2).

Concurrently, social scientists have also started inquiries into shadow banning.[6] Some have tried to detect shadow banning using computational methods, while others have investigated user experiences and perceptions. These studies tend to define shadow banning broadly, and focus on visibility remedies.[7]

What seems to unite the previous and present meanings of shadow banning, is a particular form of secrecy. Whether as account suspension or as visibility restriction, shadow banning has always referred to content moderation sanctions which the affected user is unable to detect. In this regard, shadow banning articulates a specific type of transparency critique; whereas much criticism addresses the *grounds* for content moderation, asking *why* certain items have been actioned and not others, shadow banning speaks to the prior question: *what* items have been actioned?[8] In algorithmic terms, shadow banning speaks to an opacity of content moderation's outputs, rather than logics or inputs.

Shadow banning therefore raises distinct normative concerns; secret sanctions are even more difficult to hold to account or resist than unexplained sanctions. From a due process perspective, unexplained sanctions pose a threat to foreseeability, reasoned deliberation, and legal due process, but secret sanctions do all this and more; they thwart practically all possibilities for individual and collective resistance.[9] This makes shadow banning an especially deep and controversial form of secrecy.

6 Sarah Myers West, 'Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms' (2018) 20 *New Media & Society* 4366. Cotter, "Shadowbanning is not a thing" (n 2). Erwan Le Merrer, Benoît Morgan and Gilles Trédan, 'Setting the record straighter on shadow banning' (2021) *IEEE INFOCOM 2021-IEEE Conference on Computer Communications* 1. Kokil Jaidka, Subhayan Mukerjee and Yphtach Lelkes, 'Censorship on social media: The gatekeeping functions of shadowbans in the American Twitterverse' (2022). SSRN Draft Paper <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4087843> accessed 28 June 2022. Monica Horten, 'Algorithms Patrolling Content: Where's the Harm? An empirical examination of Facebook shadow bans and their impact on users' (2021) SSRN Draft Paper <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792097> accessed 28 June 2022.

7 Le Merrer, Morgan and Tredan, 'Setting the record straighter on shadow banning' (n 6).

8 Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) *Big Data & Society* 1 <https://doi.org/10.1177/2053951719897945> accessed 19 September 2022. Nicolas Suzor and others, 'What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation' (2019) 13 *International Journal of Communication* 18.

9 Jeremy Waldron, 'The Rule of 'Law'. *Stanford Encyclopedia of Philosophy* (22 June 2016) <https://plato.stanford.edu/entries/rule-of-law/#ProcAspe> accessed 22 September 2022. Jennifer Cobbe, 'Algorithmic censorship by social platforms: Power and resistance' (2021) 34 *Philosophy & Technology* 739.

A question we must bracket for now, is how to define content moderation sanctions such as visibility remedies. In many shadow banning disputes, I will argue further below, the true disagreement may not be empirical—has the item been moderated or not?—but rather conceptual—what does it mean for an item to be moderated? In the context of algorithmic ranking and visibility management, many practices tread an unclear line between content moderation, as a process of content classification and enforcement, versus content curation, as the process through which platforms select for relevance and 'filter abundance into a collection of manageable size'.[10] I return to this problem below.

2.2 Techniques: how do platforms shadow ban?

Shadow banning is a matter of both policy and design. As a matter of policy, shadow banning is per definition an action which is not disclosed to the affected user. As a matter of design, some moderation remedies can be observed by users even when they are not disclosed—or even despite efforts to conceal them. Content moderation leaves 'traces', and some remedies leave clearer traces than others.[11] Shadow banning occurs when a traceless remedy is not disclosed. As I will argue below, the conventional methods of takedown and account suspension are relatively self-evident, leaving traces even when platforms try to conceal them, whereas visibility remedies are inherently opaque unless deliberately disclosed. This makes the policy question of notice all the more salient for these novel techniques.

Content takedown and account suspension are self-evident because they cut off engagement by all other users. Platforms may try to conceal this fact by presenting an alternative reality to the affected user, giving them the false impression that their content is still online whereas in fact nobody else can see it. But these methods are unlikely to mislead users for long, since they cause all engagement to grind to a halt (views, likes, comments, and so forth). All but the least popular users, therefore, are likely to notice that something is amiss. And since these measures cut off all engagement, content takedowns and suspensions are also relatively straightforward to test, for instance by logging off, switching to a different account, or asking a friend to check for access. For these reasons, concealed takedowns and suspensions do not afford enduring secrecy.

10 Kerstin Thorson and Chris Wells, 'Curated flows: A framework for mapping media exposure in the digital age' (2016) 26 *Communication Theory* 309.

11 Tarleton Gillespie, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 *Social Media+ Society* <https://doi.org/10.1177/205630512211175> accessed 19 September 2022.

Visibility remedies, by contrast, tend to be subtler in their effects. Their effects vary, since visibility remedies can take various forms.[12] Platforms can remove content entirely from a given feature ('delisting'), reduce its relative prominence within that feature ('demotion'), or impose some other restriction such as a disclaimer or warning label. These modalities can each apply to different recommendation (sub)systems. For instance, Twitter's arsenal of visibility remedies includes search delisting; search suggestion delisting; and reply deboosts, which demote the target's replies to the bottom of the page and hide it behind a 'show more replies' prompt.[13] In theory, visibility restrictions can also be personalised towards specific audiences; hiding an item from certain cohorts but not from others. Through these and other features, platforms conduct a complex 'management of visibilities' that steers and nudges audiences in more or less subtle ways.[14]

The problem with observing visibility remedies is, in essence, that visibility on platforms fluctuates constantly and on a personalised basis. Content visibility is governed by complex recommender and search systems, which operate through recursive interactions between user behaviour and machine-learning optimisation algorithms, which influence and alter each other over time.[15] In this dynamic, volatile process of content curation, visibility restrictions are simply one factor out of very many, and their impact on overall outcomes may be difficult or even impossible to discern.[16] And since visibility outcomes are personalised to individual users, even observing these basic outcomes at a systemic level is challenging.[17]

Visibility restrictions are at their most noticeable when they result in especially steep drops in an item's traffic or engagement.[18] But even this is not conclusive evidence. The same drops can also be attributed to user-related changes such as weakening

12 Gillespie, 'Do not recommend' (n 11). Eric Goldman, 'Content Moderation Remedies' (2021) 28 *Michigan Technology Law Review* 1.

13 Jaidka, Mukerjee and Lelkes, 'The gatekeeping functions of shadowbans in the American Twitterverse' (n 6).

14 Mikkel Flyverbom, *The Digital Prism: Transparency and managed visibilities in a datafied world* (Cambridge University Press 2019).

15 Chapter 2 above (Paddy Leerssen, 'The soap box as a black box: regulating transparency in social media recommender systems' 11(2) (2020) European Journal of Law and Technology <https://ejlt.org/index.php/ejlt/article/view/786>.)

16 Jaidka, Mukherjee and Lelkes, 'The gatekeeping functions of shadowbans in the American Twitterverse' (n 6). Le Merrer, Morgan and Tredan, 'Setting the record straighter on shadow banning' (n 6). Horten, 'Algorithms patrolling content: where's the harm?' (n 6).

17 Balazs Bodó and others, 'Tackling the Algorithmic Control Crisis: The Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents' (2018) 19 *Yale Journal of Law and Technology* 133.

18 Cotter, "Shadowbanning is not a thing" (n 2).

audience interest or intensified competition from rival uploaders.[19] The cause could also be a structural change in the platforms' ranking methods, rather than individually targeted sanctions. These competing explanations are difficult to rule out since most platforms do not disclose detailed recommendation and engagement data to their uploaders or to third parties. Most platforms offer little more than aggregate counts of views and engagements, and not the types of recommendation and analytic data that would allow users to observe shadow banning's effects, and distinguish them from more routine operations. Furthermore, allegations of shadow banning in specific cases tend to be ignored by platforms, or responded to only partially, even though they admit at a policy level that they employ visibility restrictions.[20] For Kelley Cotter, this strategy amounts to a form of 'gaslighting', which maintains the platform's 'epistemic authority' over shadow banning allegations while delegitimising valid concerns as paranoia or conspiracy theory.[21]

The platform's epistemic authority over visibility remedies is not absolute, however. Academics and other experts have been able to demonstrate undisclosed demotions. By collecting ranking data at scale, with the help of bots or user participants, one can detect especially drastic and targeted changes to recommendation trends, which permit few other explanations than a targeted restriction.[22] But these sophisticated measures are out of reach for the vast majority of users. Furthermore, as Frank Pasquale has suggested, platforms might in theory design their downranking measures adversarially to minimise the risk of detection, for instance by downranking items gradually over time rather than instantaneously.[23] Conversely, by the same logic it follows that the most restrictive *delisting* measures are relatively more self-evident than weaker forms of demotion, since their effects are more pronounced. Researchers have therefore been able to create tools which test for delisting automatically, such as Shadowban.eu

19 Gillespie, 'Do not recommend' (n 11).

20 Platform denials are often based on restrictive (and perhaps misleading) conceptions of shadow banning. One official statement by Twitter (2018) denied shadow banning by defining it as: "deliberately making someone's content *undiscoverable to everyone except the person who posted* it, unbeknownst to the original poster" (emphasis mine). This statement still fails to clarify whether visibility reductions are being applied without notice. Kelley Cotter observes a similar strategy in Instagram's communications: "while Instagram's statements avoid obvious falsehoods, they omit important clarifying information, for example a clear and consistent definition of shadowbanning". See: Cotter, "Shadow banning is not a thing" (n 2).

21 Cotter, "Shadowbanning is not a thing" (n 2).

22 Jaidka, Mukherjee and Lelkes, 'The gatekeeping functions of shadowbans in the American Twitterverse' (n 6). Le Merrer, Morgan and Tredan, 'Setting the record straighter on shadow banning' (n 6). Horten, 'Algorithms patrolling content: where's the harm?' (n 6).

23 Frank Pasquale, *The Black Box Society: The secret algorithms that control money and information* (Harvard University Press 2015).

and Whosban.eu.[24] These tools can instantly test whether specific accounts have been delisted from Twitter's search and autosuggest features by querying relevant phrases. Doing the same for demotion would be more challenging. Ironically, then, for their victims and for the public at large, visibility restrictions are often invisible.

An additional category of opaque moderation techniques is demonetisation, which renders items ineligible for advertisement revenue-sharing programs (i.e. monetisation). In a study of YouTube's demonetisation policies, Robyn Caplan and Tarleton Gillespie note that these measures can be difficult for users to observe.[25] Much like visibility restrictions, the problem stems from volatile engagement patterns combined with a lack of granular data access. Since YouTube's disbursement statements did not break down revenue for individual videos, users were usually unable to discern whether any of their videos might have been demonetised—let alone which of their videos in particular might have been actioned. In 2018, YouTube changed course and started disclosing monetisation status on a per-video basis. Once monetisation decisions came to be known by users, they quickly prompted vigorous criticism and resistance from users, who accused YouTube of inconsistency and discrimination in their policies.[26] Some users sought to hold YouTube accountable through public criticism, whilst others resisted the policy by switching to other platforms or other revenue models (e.g. direct donations).[27] This episode speaks to the importance of notice policies for unobservable remedies such as demonetisation, delisting and demotion. With notice, they encounter resistance. Without notice, they act as shadow bans.

2.3 Policies: why do platforms shadow ban?

In light of the above, it should be clear why shadow banning concerns revolve primarily around visibility remedies, and, to a lesser extent, demonetisation. The basic problem with these remedies is that, unless notified, users struggle to ascertain whether or not they have been sanctioned. Explaining the shadow banning phenomenon therefore entails two discrete questions: Why do platforms deploy visibility restrictions? And why do they refrain from notifying them?

24 Le Merrer, Morgan and Tredan, 'Setting the record straighter on shadow banning' (n 6).

25 Robyn Caplan and Tarleton Gillespie, 'Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy'. 6(2) *Social Media+ Society* <https://doi.org/10.1177/2056305120936636> accessed 15 September 2022.

26 Ibid.

27 Ibid.

To start with the first question: platforms have only recently started to intensify their use of visibility remedies, as a supplement to conventional takedown strategies.[28] In particular, visibility remedies are used to manage new controversies which fall short of violating the law, such as disinformation and political extremism, and to enforce content quality standards such as proscriptions against 'clickbait'. To justify intervention on such issues, and deflect accusations of censorship, platforms and policymakers alike have embraced visibility reduction as a more proportionate, less restrictive alternative to removal. 'We're not arguing for censorship, we're arguing just take it off the page, put it somewhere else.', Google CEO Eric Schmidt has claimed.[29] According to Facebook CEO Mark Zuckerberg, platforms ought not to become the 'arbiters of truth' in responding to disinformation, and instead 'we feel like our responsibility is to prevent hoaxes from going viral and being *widely distributed*' (emphasis mine).[30] This turn to visibility management forms part of a broader reframing of platform culpability from publication to amplification, i.e. the granting of excessive visibility.[31] Its slogan: the widely-cited adage by Renee DiResta that 'free speech is not free reach'.[32]

What seems to be missing from these accounts is the issue of transparency. Visibility remedies being less transparent, and leading to shadow bans, they are arguably *more* restrictive for users, not less so.[33] Due process and the rule of law demand that subjects are able to adapt their behaviour towards compliance and to contest wrongful decisions.[34] Imposing sanctions in secret contradicts all these principles. Without adequate disclosure, therefore, visibility management is the opposite of proportionate: the most sensitive edge-cases end up being governed through the least transparent means. Instead of a Ministry of Truth, we get a secret police.

28 Gillespie, 'Do not recommend' (n 11).

29 Matthew Wisner, 'Google's Eric Schmidt Responds to Verizon, AT&T Pulling Ads From YouTube', *Fox Business* (23 March 2017) <https://www.foxbusiness.com/features/googles-eric-schmidt-responds-to-verizon-att-pulling-ads-from-youtube> accessed 28 June 2022.

30 Kara Swisher, 'Zuckerberg: The Recode interview', *Vox* (8 October 2018) <https://www.vox.com/2018/7/18/17575156/mark-zuckerberg-interview-facebook-recode-kara-swisher> accessed 28 June 2022.

31 Daphne Keller, 'Amplification and its discontents: why regulating the reach of online content is hard', Knight First Amendment Institute (8 June 2021) < https://knightcolumbia.org/content/amplification-and-its-discontents> accessed 25 September 2022.

32 Renee Diresta, 'Free Speech Is Not the Same As Free Reach'. *Wired Magazine* (30 August 2018). https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/ accessed 28 June 2022.

33 Horten, 'Algorithms patrolling content: where's the harm?' (n 6).

34 Nick Suzor, 'Digital Constitutionalism': Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4(3) *Social Media + Society* <https://doi.org/10.1177/2056305118787812> accessed 19 September 2022. Danielle Citron and Frank Pasquale, 'The scored society: Due process for automated predictions' (2014) 89 *Washington Law Review* 1.

The question of disclosure therefore appears crucial to visibility management strategies. And yet, most platforms do not to disclose these measures to affected users. Why not? Platforms have several incentives toward secrecy. One factor may be the cost and complexity of disclosure; Gillespie notes that visibility restrictions are more complex than other remedies, and not as amenable to meaningful disclosure.[35] Though visibility remedies are certainly more complex than removal, I will argue in Section 4 below that disclosing demotions ought still to be technically feasible. There are other incentives at play as well.

Monica Horten sees shadow banning as a strategy grounded in the adversarial logics of computer security.[36] From the moderator's perspective, secret sanctions can be a convenient way of mitigating resistance from users deliberately seeking to skirt the rules, such as professional influencers or commercial spammers. For instance, users might try to 'game the algorithm' by creating new accounts or adjusting their content.[37] Still, what counts as legitimate compliance, and what amounts to illegitimate 'gaming', is decided by the platform themselves and in practice often deeply ambiguous.[38] A due process perspective might aim to clarify these standards inasmuch as possible, as a guide to user conduct. But from a security perspective such strategic ambiguity may be more effective; the path of least resistance.

Although defences of shadow banning are often cast in the technocratic, adversarial language of security and circumvention, there are also political and reputational considerations at stake. As Gillespie puts it, secretive visibility remedies can be a means to avoid public accountability.[39] Content moderation is after all a risky business —as the litany of scandals, protests, boycotts, regulatory inquiries and legal filings reminds us. From a business perspective, the safest best may often be to keep it secret. Although platforms claim to embrace transparency and accountability, their track records show the opposite; many important transparency reforms are only carried out under public pressure or legal obligation.[40] Facebook long maintained a secret

35 Gillespie, 'Do not recommend' (n 11).

36 Horten, 'Algorithms patrolling content: where's the harm?' (n 6).

37 Kelley Cotter, 'Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram' (2019) 21 *New Media & Society* 895. Caitlin Petre, Brooke Erin Duffy and Emily Hund, '"Gaming the system": Platform paternalism and the politics of algorithmic visibility' (2019) 5(4) *Social Media+ Society* < https://doi.org/10.1177/20563051198799> accessed 25 September 2022.

38 Thomas Poell, David Nieborg and Brooke Erin Duffy, *Platforms and cultural production* (John Wiley & Sons 2021).

39 Gillespie, 'Do not recommend' (n 19).

40 Monika Zalnierute, '"Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism' (2021) 8 *Critical Analysis of Law* 39.

programme known as XCheck which exempted high-profile accounts from their routine content moderation programs, with the explicit goal of avoiding errors that might lead to scandal.[41] Examples such as this illustrate clearly that platforms see their content moderation decisions as a source of reputational risk, which secrecy might serve to mitigate. Shadow banning doesn't just outwit bad actors; it also avoids bad press.

Overall, then, the incentives toward shadow banning are several. Its most important driver may be the general turn to visibility remedies, as a response to disinformation and other recent controversies around 'lawful but awful' content. These new techniques are less observable, and therefore lend newfound significance to transparency safeguards; safeguards not just for the reasons behind moderation processes, but their basic outcomes. Platforms have several reasons not to offer these safeguards: from implementation costs to anti-circumvention considerations to their general avoidance of public accountability. All these factors suggest that shadow banning will likely persist unless platforms face sufficient pressure to end it. Enter: the Digital Services Act.

3. Transparency rules for content moderation in the Digital Services Act

The Digital Services Act (DSA) is not the first legislation to regulate transparency in content moderation, but it is the first to introduce a general right against shadow banning.[42] The following section proceeds by introducing the general features of the DSA's notice-and-action framework for content moderation, including its definition of shadow banning in Recital 28. It then highlights two key provisions that regulate shadow banning practice: Article 14 on Terms of Service, and Article 17 on the Statement of Reasons.

41 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/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353> (accessed 28 June 2022). More generally, see: Caplan and Gillespie, 'Tiered Governance' (n 25).

42 The platform-to-business regulation or 'P2B Regulation' contains a similar set of rights in Articles 3 and 4. This instruments was adopted in 2019, only three years before the DSA. This contribution focuses on the DSA since its rights are both deeper in substance and broader in scope; many of the DSA's safeguards apply to all users of hosting services, including platforms, whereas the P2B Regulation applies only to business users of online intermediation services (See P2B Regulation, Articles 3 and 4). Most relevant due process issues covered by the P2B Regulation are therefore covered by the DSA's rules as well, whereas the inverse is not true. Some additional comparative reflections are included at the end of this section.

3.1 The DSA's notice-and-action framework for content moderation

The DSA is a lengthy and complex piece of legislation, but it is fair to say that its main concern is content moderation. This it regulates in three ways. First, it restates, with only minor revisions, the pre-existing 'safe harbour' regime governing internet services' liability for unlawful user-generated content.[43] Second, it outlines a comprehensive due process framework for all content moderation actions, known as the 'Notice-and-Action' framework. Third, the DSA sets out due diligence obligations for the very largest platforms, known as 'Systemic Risk Mitigation'. Most relevant for our purposes is the second element: notice-and-action due process.

The DSA's notice-and-action framework applies to all content moderation actions—a concept which it defines broadly. Whereas earlier content moderation laws have concerned themselves almost exclusively with content removal and account suspension, in what Eric Goldman has termed the 'binary leave up / take down paradigm', the DSA innovates with an expansive understanding of content moderation actions that includes non-removal remedies.[44] Content moderation is explicitly defined to include not just takedown or account suspension, but also demonetisation and visibility restrictions.[45] Recital 55 defines visibility remedies in greater detail, and even mentions shadow banning explicitly:

> *Restriction of visibility may consist in demotion in ranking or in recommender systems, as well as in limiting accessibility by one or more recipients of the service or blocking the user from an online community without the user knowing it ('shadow banning').*

This recital clearly uses shadow banning in the original, narrow sense of secret account suspensions, rather than the modern, broad sense of secret visibility restrictions. From a legal standpoint this matters little, however, since the phrase 'shadow banning' is only used in this recital and does not return in the DSA's actual enacting provisions (i.e. its 'articles'). Going forward, lawyers would do well to keep in mind this gap between

43 Its predecessor is Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('E-Commerce Directive').

44 Eric Goldman, 'Content Moderation Remedies' (2021) 28 *Michigan Technology Law Review* 1. See also: Wolfgang Schulz and Stephan Dreyer, Governance von Informations-Intermediären - Herausforderungen und Lösungsansätze - Bericht an das BAKOM (Hans Bredow Institut Research Report 2020) <https://leibniz-hbi.de/de/publikationen/governance-von-informations-intermediaeren-herausforderungen-und-loesungsansaetze> accessed 26 September 2022. Martin Husovec and Irene Roche Laguna, 'Digital Services Act: A Short Primer', in: Husovec and Roche Laguna, *Principles of the Digital Services Act* (Oxford University Press forthcoming 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4153796> accessed 25 September 2022.

45 DSA, Article 3(t).

statutory and popular usage. But regardless of these semantics, the fact remains that visibility remedies, which attract the bulk of shadow banning speculation, are indeed recognised as content moderation actions, and are therefore equally subject to the DSA's notice-and-action procedures.

The DSA's notice-and-action framework regulates platforms as follows. (Technically, most of these rules apply not just to platforms but to all 'hosting services' which store user-generated content, but for ease of use I limit my discussion to platforms.) Its cornerstone is Article 14 DSA on Terms and Conditions, which lays down two key principles. First, the rules governing online services' content moderation must be published in their Terms and Conditions, in 'clear and unambiguous language'. Second, these rules must be enforced 'in a diligent, objective and proportionate manner', and with due regard to the interests and fundamental rights involved.[46] Article 16 adds that platforms must offer a notice mechanism through which third parties can flag content for content moderation review. Crucially for our purposes, Article 17 then requires that platforms provide a Statement of Reasons to the affected uploader for each content moderation action, whether made in response to a notice or by the service's own initiative. These decisions must also be open to appeals through internal complaint handling (Article 20) and through out-of-court dispute settlement (Article 21). Taken together, this framework reflects basic principles of due process: every sanction—i.e. any deprivation of lawful interests—must be governed by clear and foreseeable rules; must be notified and explained to the affected users; and must be open to appeals.[47] As we will see below, this leaves little room for shadow banning.

This is only a basic sketch. The DSA introduces many more transparency rules besides, but these are generally less relevant to the issue of shadow banning. For instance, the DSA also contains public reporting requirements for content moderation actions (e.g. Articles 15, 23, and 42), explanation duties for recommender systems (Article 27), and data access for regulators and researchers (Article 40). Yet these provisions are not designed to shed light on individual cases, and therefore shed light no on shadow bans. Below I focus on Article 14's Terms and Conditions and Article 17's Statement of Reasons.

3.2 Article 14 DSA on Terms and Conditions

Article 14(1) DSA demands that platforms codify their content moderation rules. In 'clear and unambiguous language', their Terms and Conditions must set out

46 Naomi Appelman, João Quintais and Ronan Fahy, 'Using Terms and Conditions to apply Fundamental Rights to Content Moderation: Is Article 12 DSA a Paper Tiger?' (Verfassungsblog 1 September 2021) <https://verfassungsblog.de/power-dsa-dma-06/> accessed 16 September 2022.

47 Suzor, 'Digital Constitutionalism' (n 34).

information about the restrictions they impose regarding user-generated content. This disclosure 'shall include information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review.' I refer to this as the codification principle, since it reflects the rule of law principles of legality, foreseeability, and accessibility.[48]

This codification principle is relatively novel in platform regulation. In keeping with the binary approach focused on unlawful content, most content moderation laws have left platforms' internal rules largely unregulated.[49] Precursors to the DSA's codification principle can already be found in national court precedents, which have found overly vague Terms to be incompatible with fundamental rights, consumer protection, and general principles of contract law.[50] In this light, Article 14 DSA's codification principle is not strictly new, but instead serves to clarify, and perhaps strengthen, a pre-existing duty for platforms to stipulate clear and specific content moderation policies in their service contracts.

Most major platforms already publish content policies, and these have become more detailed over time. Still, these voluntary efforts continue to be criticised for their lack of detail, and Article 14 DSA might force further reforms by holding them to its standard of clear and unambiguous language. Its impact may be especially significant for non-takedown remedies, which tend to be given short thrift in platforms' current Terms. Non-takedown policies are more often discretionary, lacking any clearly formulated policy or with relatively generic policies such as restrictions on 'inappropriate' or 'borderline' content.[51] Facebook recently published a systematic overview of its (down)ranking policies, known as its 'Content Distribution Guidelines', but this is the exception to the general rule that the policies for most non-takedown

48 Ibid.

49 Some exceptions: variants on the DSA's codification principle—though far more narrowly tailored in scope—can be found in recent sectoral frameworks such as the recent Platform-to-Business Regulation (as regards the ranking of business users by ecommerce platforms) and the Audiovisual Media Services Directive (as regards the protection of minors, hate speech and terrorism on video sharing platforms). See: P2B Regulation, Article 3. AVMS Directive, Article 28(b)(3)(a).

50 Mattias Kettemann and Anna Sophia Tiedeke, 'Back up: Can users sue platforms to reinstate deleted content?' (2020) 9(2) *Internet Policy Review* <https://doi.org/10.14763/2020.2.1484> accessed 19 September 2022.

51 Amélie Heldt, 'Borderline speech: caught in a free speech limbo?', *Internet Policy Review* (15 October 2021) <https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510> accessed 15 September 2020. Horten, 'Algorithms patrolling content: where's the harm?' (n 6).

remedies are not published in the same systematic detail as for takedown.[52] Article 14 DSA would demand a more systematic and comprehensive approach to such documentation for all major platforms and for all moderation measures, and thus help to shed light on visibility management policies.

Still, Article 14 DSA is only a partial solution for shadow banning. If enforced properly, it might provide some reassurances by improving the foreseeability of platform policies and helping users to self-assess their compliance. For this to succeed, the Terms would need to be both detailed and clear, and even then, it would only help relatively sophisticated and proactive users—those with the wherewithal to seek out and study these guidelines. Most users, we know, do not consult Terms and Conditions, though experts do.[53] Even towards experts, there is little cause for optimism about the foreseeability that Terms can provide. Like all contracts, platform Terms face the basic problems of indeterminacy and contractual incompleteness; no statute or contract is ever sufficiently detailed to cover all contingencies, and will inevitably leave room for interpretation. Even legal doctrines with centuries of jurisprudence behind them, such as defamation or fair use, continue to divide lawyers, leaving little hope that enforcement of platform Terms should ever be any more foreseeable. Indeed, excessively detailed codifications may not even be desirable due to trade-offs with flexibility and substantive fairness, which could unduly hamper moderators in unforeseen circumstances.

Adding to this challenge of foreseeability are the practical constraints of content moderation at scale. Content moderation is not a judicial process of careful legal reasoning, but an industrial process that occurs at massive scales through standardised routines and procedures.[54] In light of its massive scale, Evelyn Douek proposes that content moderation is best understood as an administrative bureaucracy, rather than as a judiciary carefully weighing individual cases.[55] And even this administrative analogy, as Douek herself acknowledges, may overstate the degree the role of human judgement. Human moderators, if at all involved, are typically forced to decide on moderation actions through snap judgements and crude heuristics, and rarely have

52 Facebook, 'Sharing Our Content Distribution Guidelines', *Facebook Newsroom* (23 September 2021) <https://perma.cc/BRT3-7XC8> accessed 28 June 2022.

53 Archon Fung, 'Infotopia: Unleashing the Democratic power of transparency' (2013) 41 *Politics & Society* 183.

54 Sarah Roberts, *Behind the Screen: Content moderation in the shadows of social media* (Yale University Press 2021).

55 Evelyn Douek, 'Content Moderation As Administration' (2022) 136 *Harvard Law Review*, forthcoming. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4005326> accessed 15 September 2022.

time for careful deliberation or fact-finding.[56] For instance, Facebook instructed its moderators to classify content as terrorist propaganda for the mere *mentioning* of certain terrorist organisations.[57] Many more decisions are automated entirely.[58] Based on automated machine-learning classifiers, these automated decisions operate through statistical inferences which may be even further removed from content policies as expressed in human language.[59] In short, the true drivers of content moderation, as operationalised in everyday practice, are often far removed from the policy principles which they nominally serve to enforce.

For all these reasons, Article 14's Terms and Conditions contribution to foreseeability is likely to be modest. Its most important function may be not as an *ex ante* guide to user conduct, nor as a factual or instrumental description of content moderation logics, but rather as an *ex post* rubric for appeals and error correction; a form of *justificatory* transparency which aims to establish and vindicate individual rights.[60] Of course, the problem with shadow banning is that it precludes all such occasions for appeal and error correction; to even begin contesting these decisions, they must be made known to the affected users. For that, we must turn to Article 17.

3.3 Article 17 DSA on the Statement of Reasons

Article 17 DSA demands that each moderation action be accompanied by a 'Statement of Reasons' to the affected user. This Statement must include the following information: (1) the measure taken; (2) the legal or contractual violation that this measure responds to; (3) the facts and circumstances relied on in taking the decision; (4) information on the role of automated decision-making in this action, (5) whether or not the measure was taken in response to a third party notice; and (6) the user's possibilities for redress.[61]

Article 17 fulfils at least two distinct functions: notification and explanation. Notification makes users aware of sanctions, whereas explanation aims to give reasons

56 Roberts, *Behind the screen* (n 54).

57 Sam Biddle, 'Revealed: Facebook's Secret Blacklist of "Dangerous Individuals and Organizations'. *The Intercept* (12 October 2021). <https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/> accessed 26 June 2022.

58 Roberts, *Behind the screen* (n 54). Hannah Bloch-Wehba, 'Automation in moderation' (2020) 52 *Cornell International Law Journal* 41.

59 Amélie Heldt cites the telling example of Facebook misclassifying a pair of onions as due to the 'overtly sexual manner' they were positioned—evidently the result of a machine-learning classification error. See: Heldt, 'Borderline speech'(n 51).

60 Margot Kaminski, 'Understanding Transparency in Algorithmic Accountability' in: Woodrow Barfield (ed.), *Cambridge Handbook of the Law of Algorithms* (Cambridge University Press 2020).

61 DSA, Article 17(3).

for those sanctions. Explanation is a crucial feature of due process, and raises many difficult policy questions in the context of (automated) content moderation.[62] But shadow banning, as discussed, is primarily a problem of notification; it speaks to an opacity of decisions moreso than reasons. For shadow banning, Article 17 DSA could be highly impactful even if it offered no explanation at all. Indeed, one might say that Article 17 DSA's notification duty amounts to a prohibition on shadow banning, which is characterised by a lack of notification.

Article 17 DSA's notification duty does contain exceptions, however. First, it does not apply to moderation actions taken in response to removal orders by public authorities, as regulated under Article 9 DSA. This exception is not fully relevant to the problem of shadow banning, since it is geared toward takedown actions and not visibility restrictions. Second, and more importantly for our purposes, Article 17(1) DSA exempts content moderation actions affecting 'deceptive high-volume commercial content'. This exception is worth discussing in detail, as it is here that the DSA attempts to balance the competing interests at stake in shadow banning.[63]

What this rule seems to envision, is a narrow exception permitting shadow banning in the context of advertising spam—i.e. 'high-volume commercial content'. That the EU legislator should side with secrecy here stands to reason; advertising spam is perpetrated by relatively persistent and well-resourced adversaries, and appeals to no significant public interests. In advertising spam, therefore, the public interest in transparency and due process is relatively low, whereas the public interest in secrecy (and thus the effective combating of adversarial actors) is relatively high. A broader exemption might also have included political spam, in what is known as 'information operations' or 'coordinated inauthentic behaviour'.[64] But the DSA's focus on commercial content suggests that such political activity is too sensitive from a public interest perspective to permit unaccountable shadow banning.

62 In particular, machine-learning complicates most attempts to explain what factual circumstances are involved in a decision. See: Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) *Big Data & Society* 1 <https://doi.org/10.1177/2053951719897945> accessed 19 September 2022.

63 The DSA's legislative history supports this reading; in the original proposal, the Statement of Reasons applied only to takedown decisions, and did not contain any exemptions for commercial content. The same round of amendments which then expanded this rule to cover all moderation actions, also added the exemption for high-volume commercial content.

64 Fabio Giglietto and others, 'It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections' (2020) 23 *Information, Communication & Society* 867.

What is surprising is the proviso that this commercial content must be 'deceptive' for shadow banning to be permitted. This is a substantial narrowing. After all, spam is typically unwelcome regardless of whether it is true. And for platforms to check for truthfulness is a major burden in practice, as they moderate millions of such items every day, and truth is difficult to assess at scale. It is also unclear how this exception will apply to moderation actions taken against *users*, given that the exception refers to deceptive commercial *content*. Overall, then, the exception is so narrow as to risk being almost unusable in practice, leaving little room for shadow banning under the DSA at all.

The DSA's secrecy rules can be contrasted with those in the P2B Regulation, which contains comparable transparency rights for business users. Here the exceptions are generally broader and more flexible. First, the P2B Regulation's statement of reasons in Article 4 takes an actor-based approach, and simply permits secrecy in cases where the business user in question 'has repeatedly infringed the applicable terms and conditions'.[65] This actor-based approach will likely appeal to platforms since it is far more practicable to assess repeat violations than to assess veracity. Yet the concept 'repeat infringement' might threaten due process for ordinary users if it is interpreted too broadly.[66] Second, the P2B Regulation's disclosure rules for ranking in Article 5 attempt to manage security and circumvention concerns by introducing an exemption disclosure of 'any information that, with reasonable certainty, would result in the enabling of deception of consumers or consumer harm through the manipulation of search results.'[67] This exception based on the substance of disclosures seems to enable platforms to modulate the level of detail given in explanations, without necessarily impinging on the basic, prior safeguard of notification. In sum, whereas the DSA's secrecy rules focus on the nature of the moderated *content*, the P2B shows how considering the *actors* and *disclosures* might also be relevant parameters for the balancing of transparency against secrecy.

In this light, the DSA's shadow banning exceptions are not only narrow but somewhat inflexible, in that they focus only on the moderated content and do not permit other factors to be taken into account. In addition to the P2B Regulation's actor- and

65 P2B Regulation, Article 4(5).

66 For instance, YouTube's 'copyright strikes' systems escalates sanctions users as little as three violations over as long as a six-month period. This approach has been criticised for the risk of chilling effects on user activity. Annemarie Bridy, 'The Price of Closing the Value Gap: How the Music Industry Hacked EU Copyright Reform' 22 *Vanderbilt Journal of Entertainment and Technology Law* 323 (2020), citing Shoshana Wodinsky, 'YouTube's Copyright Strikes Have Become a Tool for Extortion', *The Verge* (11 February 2019) <https://www.theverge.com/2019/2/11/18220032/youtube-copystrike-blackmail-three-strikes-copyright-violation>.

67 P2B Regulation, Article 5(6).

information-based exceptions, another important factor that might in theory be considered is the nature of the *enforced rule*. For instance, actions against child sexual abuse imagery or cyberstalking might justify a greater degree of secrecy than those against clickbait or conspiracy theories. As to account-based approaches, combatting spam might also benefit from an exception not just for repeat infringements but also for *new* accounts; rapidly creating new accounts is an important strategy for spammers to circumvent account suspensions and terminations. But at present, a brand new account with zero followers or post history seems to be entitled to the same due process treatment as established users. Clearly, the cost-benefit analysis for due process is complex and may be vary significantly across all these different cases—from a transaction cost perspective, from a security perspective and from a public interest perspective. But for Article 17 DSA, all that matters is whether the item contains high-volume deceptive commercial content.

More fundamentally, the DSA's design is inflexible in that it bundles all relevant due process rights—notice, explanation and appeals—into the singular concept of a 'moderation action'. These actions receive the full suite of safeguards, or, in the case of deceptive commercial spam, none at all. In practice there may be a large set of edge-cases where integral explanation and/or appeal could be onerous, or too sensitive from a security perspective, but where a bare notice right could still be of substantial value as a bulwark against shadow banning and as a minimal precondition for legal and social accountability. In this light, the DSA's attempt at balancing is somewhat rudimentary, and in future may benefit from further refinement, such as by incorporating more factors into the shadow banning calculus and unbundling notice safeguards from other aspects of due process.

At present, the DSA's rigid design still deserves praise for erring on the side of transparency rather than secrecy. Current shadow banning practices have until now rested primarily on untested technocratic arguments about circumvention. As I have argued in Section 2.3 above, these claims are not only difficult for outsiders to assess but also risk giving cover to the platforms' more general disinterest in accountability and due process. The DSA, by erring on the side of transparency, will put these arguments to the test, forcing platforms to demonstrate the practical need for shadow banning (if any) and make these claims available for public scrutiny. If greater secrecy is deemed necessary in future lawmaking, then this will at least be a secrecy arrived at through public rulemaking, rather than, as present, a secrecy taken on faith from self-interested platforms.

Regardless of such future revisions, a more pressing problem for the regulation of shadow banning is how the DSA defines moderation actions in the first place; what it means for an item to moderated. As I will discuss below, the category of ranking or 'demotion' actions is especially problematic.

4. Ranking due process between moderation and curation

4.1 Defining demotion and the problem of counterfactuals

Compared to most content moderation remedies, it is not so clear what it means to 'demote' an item. Most other remedies can be summed up in relatively straightforward binaries: an item can be either left up or taken down; listed or delisted; monetised or demonetised; an account active or suspended. But when is an item 'demoted' or 'downranked', as opposed to merely 'ranked'? The basic problem here is that ranking is a zero-sum, relative process in which all items receive differential treatment, leaving no clear baseline of ordinary or default treatment for comparison. In other words, demotion lacks a clear counterfactual.

Several commentaries have already remarked on this problem of counterfactuals as a potential barrier to regulation of ranking due process and of 'demotion'. For Rachel Griffin, it counsels against a human rights approach to ranking governance.[68] Griffin argues that ranking interventions are 'difficult to frame as a clear-cut rights violation', since, after all, '[w]hat level of algorithmic visibility does anyone have a right to?'[69] Gillespie concludes that 'it is nearly impossible to be transparent about reduction policies', since, after all, '[h]ow does one measure or document reduction: what should the reduced visibility of a piece of content be compared to?'[70] Very similar objections have also been raised against the regulation of 'amplification', which refers to excessive visibility rather than restricted visibility and, in this sense, can be seen as the mirror image to demotion. Daphne Keller objects that proposals to regulate amplification are 'hard to assess, because it is hard to define', and for Luke Thorburn, Jonathan Stray and Priyanjana Bengani the concept of amplification is 'not precise enough to be used in law'.[71] This problem of counterfactuals

68 Rachel Griffin, 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality', SSRN Draft Paper (2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064738> accessed 19 September 2022.

69 Ibid.

70 Gillespie, 'Do not recommend' (n 11).

71 Daphne Keller, 'Amplification and its discontents: why regulating the reach of online content is hard', Knight First Amendment Institute (8 June 2021) <https://knightcolumbia.org/content/amplification-and-its-discontents> accessed 25 September 2022. Luke Thorburne, Jonathan Stray and Priyanjana Bengagina, 'What Will "Amplification" Mean in Court?' *Tech Policy Press* (19 May 2022) <https://techpolicy.press/what-will-amplification-mean-in-court/> accessed 19 September 2022.

in ranking regulation poses a conceptual challenge for the DSA's regulation of demotion. It also speaks to an ambiguity in the shadow banning imaginary itself. Both necessitate some underlying distinction between ordinary ranking routines and exceptional ranking sanctions.

I want to offer a slightly more optimistic and constructive account. Without denying the problem of counterfactuals, I argue that a workable legal concept of 'demotion' might still be devised through detailed engagement with specific ranking architectures. Demotion practices come into view more clearly when one recognises that the platform ranking process does not consist of one single, monolithic Algorithm, but is instead comprised of many fragmentary organisational and computational units all working in concert but fulfilling distinct functions.[72] In these complex assemblages, it is possible to distinguish certain subsystems that ascribe relevance scores to content (typically optimised for user engagement), and others that impose *ex post* maluses or bonuses on these scores based on ulterior optimisation goals, such as clickbait or hate speech classifiers. In other words, certain subsystems *produce* algorithmic relevance scores, whereas other subsystems serve only to *reduce* them.[73] The former optimises for engagement, the latter for compliance. It is these reduction decisions that most clearly constitute moderation actions. This interpretation manages the problem of counterfactuals by taking engagement optimisation as its baseline treatment, against which reductions can then be defined.

Facebook's own description of its Newsfeed ranking process (Figure 1 below) can serve as an example. It includes three main steps involving different sets of machine learning classifiers: (1) inventory or candidate generation, which selects several hundreds of possibly relevant candidates out of the pool of available content, (2) relevance scoring, which attributes initial ranking scores to all candidates based on a 'multitask model' for engagement optimisation, and (3) integrity processes, which test items for compliance with rules such as those on borderline content and spam. Whereas steps (1) and (2) appear to optimise for relevance, and together produce relevance scores, step (3) optimises for entirely different 'integrity' classifiers, often content-related, and *reduce* relevance scores. These integrity processes, then, might result in 'demotions' for purposes of EU law.

72 Bernhard Rieder and Jeanette Hofmann, 'Towards Platform Observability'(2020) 9(4) *Internet Policy Review* <https://doi.org/10.14763/2020.4.1535> accessed 19 September 2022. Nick Seaver, 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems' (2017) 4(2) *Big Data & Society*.

73 Gillespie, 'Do not recommend' (n 14).

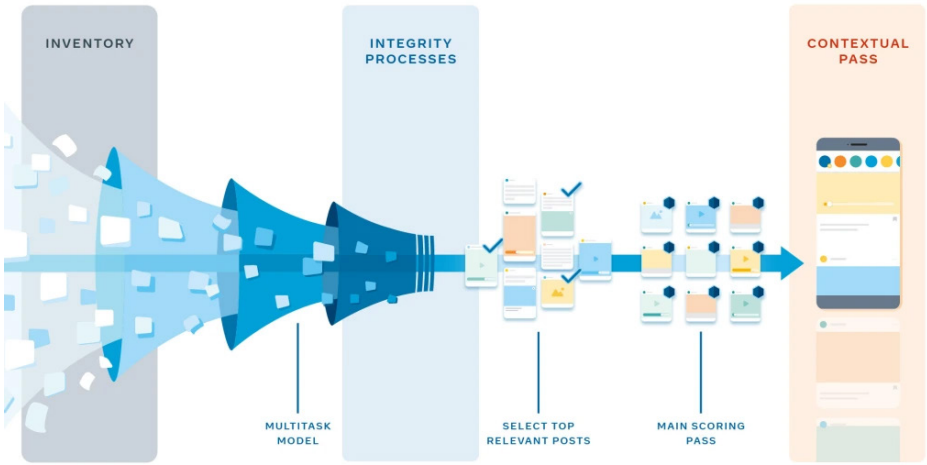

Figure 3: Facebook's schematisation of its Newsfeed Ranking Procedure[74]

Both platforms and lawmakers, therefore, appear to be coalescing around a concept of demotion as an *ex post* reduction of engagement scores. This approach is open to critique, however, and models such as Facebook's focused on 'integrity processes' risk concealing other interventions in the system and other exercises of ranking power, insofar as it does not account for the ways in which platforms regulate content throughout the relevance scoring process itself. Tarleton Gillespie warns that a regulatory focus on visibility reductions runs the risk of normalising or depoliticising the relevance scoring process itself: 'When we are fighting about particular dynamics of virality, we are not asking whether there are other logics of circulation that we should prefer'.[75] Further to this point, it is worth noting that relevance scoring is not a fixed or objective process but one that is itself iterative and politically strategic. Platforms act as gatekeepers not just by ruling on exceptions to the ranking game, but by constantly re-writing those rules over time. Optimisation goals such as 'engagement', 'relevance, or 'quality' may seem objective, but in practice their measurement entails a complex

74 This is a screenshot taken from Facebook's official website. The 'contextual pass' mentioned in this schema refers to an additional step accounting for contextual considerations such as content diversity. Akos Lada, Meihong Wang and Tak Yan, 'How Does News Feed Predict What You Want to See?' *Meta Newsroom* (26 January 2021) <https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/> accessed 28 June 2022.

75 Gillespie, 'Do not recommend' (n 14).

and value-laden weighing of competing interests.[76] Relevance scoring may therefore harbour its own forms of content regulation, which 'demotion' safeguards would then fail to capture. In this sense, a regulatory project focused on 'demotion' risks overlooking the structural or constitutive role of platform ranking power.[77] Put differently, this due process approach highlights exceptional cases of moderation at the cost of normalising more routine forms of curation.

An example to illustrate this point is the history of Facebook's reaction feature, as reported by the Washington Post.[78] The 'Like'-button has long been an important component of Facebook's engagement optimisation metrics, but in 2016 the platform added several new options including a 'Haha', 'Wow', and 'Angry' react. In order to encourage users to experiment with these new and unfamiliar features, Facebook initially measured these new reacts as a stronger form of engagement than a conventional 'Like'. Later, the platform observed that the 'Angry' emoji correlated strongly with low-quality content and disinformation. To slow the spread of this content, the platform reduced the engagement signal of Angry reacts to zero. In this way, Facebook suppressed content not by reducing its relevance scores, but instead by changing how they define relevance in the first place.

Regulating 'demotion' only as visibility *reduction*, rather than visibility *production*, fails to account for these constitutive forms of ranking power. Still, it has the advantage of constraining relatively fine-grained and targeted interventions. Even if systemic interventions like Facebook's reaction update can be implemented with specific (sets of) targets in mind, they do so indirectly based on observed patterns of user engagement rather than through directly targeted decisions. Ex post demotion interventions, by contrast, afford a relatively fine-grained form of control. They permit platforms to regulate content not only by tweaking relevance metrics but on wholly separate criteria, including automated but also manual human intervention. From a freedom of expression perspective, therefore, *ex post* sanctions may arguably raise heightened concerns of censorship or viewpoint discrimination; they provide a venue for platforms to exercise

76 Gillespie, 'The relevance of algorithms' (n 14). Natali Helberger, 'On the democratic role of news recommenders' (2019) 7 *Digital Journalism* 993. Philip Napoli, *Social media and the public interest: Media regulation in the disinformation age* (2019 Columbia University Press). Elizabeth van Couvering, 'Is relevance relevant? Market, science, and war: Discourses of search engine quality' (2007) 12 *Journal of Computer-Mediated Communication* 866. Joris van Hoboken, *Search Engine Freedom: On the Implications of the Right to Freedom of Expression for the Legal Governance of Web Search Engines* (Kluwer International 2012).

77 Griffin, 'Rethinking rights in social media governance' (n 68). Cotter, 'Playing the visibility game' (n 37).

78 Jeremy Merrill and Will Oremus, 'Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation'. *The Washington Post* (26 October 2021) <https://washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/> accessed 19 September 2022.

content-specific 'opinion power' in ways that the engagement optimisation process itself may not.[79] In this sense, ex post restrictions raise distinct risks from a fundamental rights and due process perspective, which arguably require distinct safeguards.

In light of the above, I argue that it is not entirely futile or incoherent to regulate demotion as a category of content moderation sanctions under the DSA. It is, however, technically complex and normatively incomplete. Especially in light of regulatory agencies' limited technical capacities, this complexity may provide platforms with occasions for obfuscation. Transparency is in practice performative, and it alters the practices it documents.[80] In the same way that public meeting rules push lawmakers into backchannels, Article 17 DSA might encourage platforms to hide their most controversial measures away from those sites which the law recognises as content moderation. For these reasons, platforms' descriptions of their ranking process should not be taken at face value. Rather, in order to determine the mechanisms of 'demotion', regulators must take full and independent stock of platform ranking procedures.

Even if Article 17 DSA is enforced rigorously, and all demotion is disclosed dutifully, what it probably cannot do is put an end to shadow banning *suspicions* and *allegations*. Users will continue to face sudden and inexplicable drops in visibility, if not due to targeted *ex post* reductions then due to more systemic *ex ante* adjustments to the ranking system; or simply by the ever-shifting whims of audience taste and attention. Such precarity is a structural feature of social media ranking.[81] That the law does not recognise users' rise and fall as 'content moderation', may then be of little reassurance to them. From the user perspective, these fluctuations may be functionally indistinguishable from shadow banning, and will likely continue to arouse suspicions of foul play. Helping publics to grapple with these more constitutive dimensions of ranking power demands that we move past narrow concerns with shadow banning, and the language of content moderation which it draws on, and towards a more comprehensive reckoning with the precarities of content curation.

4.2 Ranking transparency beyond the downrank: from moderation to curation

The above has shown that important aspects of ranking governance cannot be broken down into individual acts of content moderation—into discrete demotion sanctions affecting specific targets. These structural or constitutive features of content ranking are integrated into the engagement optimisation process; they *produce* ranking rather

79 Natali Helberger, 'The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power' (2022) 8 *Digital Journalism* 842.

80 Flyverbom, *The Digital Prism* (n 14).

81 Brooke Erin Duffy, 'Algorithmic precarity in cultural work' (2020) 5 *Communication and the Public* 1.

than merely reducing it. And yet, as we have seen, these structural features of ranking are themselves an important site of opinion power. Addressing these demands a more expansive approach to transparency and accountability in ranking systems, not only as an occasional site of content moderation but as structural site of content curation. What new models of disclosure come into view when we look beyond the restrictive categories of content moderation, demotion, and shadow banning?

Transparency in ranking curation can take several forms. The first step, already taken in the DSA, is Article 27's codification principle for ranking policies. Whereas Article 14, discussed previously, demands codification for content moderation actions, Article 27 DSA requires transparency of recommender systems in a general sense: platforms must disclose 'the main parameters used in their recommender systems', including 'the criteria which are most significant in determining the information suggested to the recipient of the service.' In addition, Article 27 DSA requires that platforms state 'the reasons for the relative importance of those parameters.' Since this provision is not limited to moderation sanctions, it can start to shed light on curation in a more comprehensive sense.

General codification rules such as these still face important limitations, however. Abstract descriptions of recommender algorithms struggle to shed meaningful light on their operation in practice, due to the extreme complexity of their machine-learning algorithms as well as the contingency of their interaction with user content and audiences.[82] Furthermore, Article 27 does not even require in-depth explanations, but instead merely a description of 'main parameters'. At worst, these descriptions could be so generic as to offer no practical guidance. But assuming a more robust implementation, it might function as a useful complement to individual content moderation transparency; when users experience a sudden drop in traffic, and receive no notice of individual moderation actions under Article 17 DSA, this might then prompt them to check for general updates to curation policies under Article 27 DSA.

More ambitious reforms would focus on access to ranking data. Ultimately, user concerns about shadow banning are fuelled by a lack of granular traffic data, which hinders them in observing their performance in ranking systems. The available data is often limited to view and engagement aggregates, with little information offered on actual recommendation trends and audience discovery pathways—or reserved only for paying customers.[83] Better access to this data could serve a dual function. First, access to analytic data could help to detect undisclosed instances of demotion,

82 Chapter 2 above (Leerssen, 'The soap box as a black box').

83 Some platforms are relatively generous with this data (e.g. YouTube, Instagram), whereas others choose not to disclose it (e.g. Facebook, TikTok).

and thus help to enforce Article 17 DSA's protections against shadow banning. Without this data, shadow banning will remain an 'known unknown' factor, and its enforcement could be difficult. Second, access to analytic data could help users and publics to understand curation trends in a broader sense. For instance, a users' ill-founded shadow banning concerns might be put to rest if she could observe a drop in audience engagement rates. In this way, observing ranking outcomes could be a first step towards understanding ranking conduct.[84] Ideally, such data would also be made available not only to uploaders themselves but also to other stakeholders in platform governance, and to the public at large, to support collective processes of legal and social accountability. [85]

5. Conclusion

Visibility remedies are making content moderation more nuanced, but less transparent. The blunt instruments of content takedown and account suspension were largely self-evident in their effects. But visibility remedies leave barely any trace, since they play out through dynamic and volatile ranking systems which obfuscate their effects. Recent allegations of shadow banning can be understood as a justified response to these new moderation techniques, expressing a newfound urgency around transparency safeguards for these opaque forms of content moderation.

The DSA, with its comprehensive framework for content moderation due process, makes shadow banning a legal problem. I have shown how the DSA's actual reference to 'shadow banning' risks being misinterpreted, as it refers solely to secret account suspensions and not to the broader array of secret visibility remedies. Nonetheless, the DSA's notice rights amount to a general prohibition on shadow banning in all forms, with only narrow exceptions for high-volume deceptive commercial content. This approach leaves relatively little flexibility to balance the competing interests at stake in content moderation secrecy, particularly as regards non-deceptive and non-commercial forms of high-volume spam. I have argued that future revisions may require a more nuanced set of exceptions, based not only on the affected content but also taking into account other factors such as the actors and norms at issue. Unbundling the due process rights of notice, explanation and appeal may also help in striking this balance. Although the DSA may lack nuance on these points, its choice to err on the side of transparency appears to a sensible one, since it helps to bring these balancing considerations out in the open. The case for transparency is already clear,

84 Rieder and Hofmann, 'Towards platform observability' (n 72).

85 Chapter 2 above (Leerssen, 'The soap box as a black box').

but the case for shadow banning remains speculative and undependable—tempting for platforms to exaggerate and difficult for outsiders to assess. By erring on the side of transparency, the DSA places the onus on platforms to demonstrate the practical importance of shadow banning (if any!) and make it available for public scrutiny. Should future lawmaking then opt for a return to shadow banning, then this secrecy will at least be arrived at through public rulemaking, rather than, as present, a secrecy taken on faith from self-interested platforms.

The final section of this paper has highlighted a more fundamental problem: the meaning of 'demotion' as a category of moderation sanctions. This concept is central to the shadow banning imaginary and the DSA's response to it, and yet its meaning is far from straightforward. In an attempt to refine earlier criticism, I have argued that demotion is not necessarily incoherent as a legal concept, if understood as an *ex post* modification to content relevance scores. Understood in this way, safeguards against demotion may help to shed light on relatively fine-grained and targeted exercises of platform opinion power. Still, it should be kept in mind that such demotion safeguards do not account for more the constitutive aspects of ranking governance; how platforms govern visibility not just by ruling on ranking exceptions, but by writing and constantly revising the rules of the ranking game.

5

CHAPTER 6

Seeing what others are seeing: Regulating social media for and with observability

Abstract

Algorithmic transparency is high on the agenda for social media regulation. But recent work in Science and Technology Studies (STS) questions whether this endeavour of 'opening the black box' is feasible or even meaningful, due to the sociotechnical contingency of platform behaviour. Bernhard Rieder and Jeannette Hoffman have therefore proposed a move from algorithmic transparency to platform observability; a pragmatic program aimed at securing structural, real-time access to the means of platform knowledge production. Taking a legal perspective, this paper examines the data access provisions of the EU's new Digital Services Act (DSA), and how these can be understood as an early attempt to surpass the algorithmic explanation paradigm and to start regulating for and with observability. In doing so, however, the DSA surfaces important challenges for observability regulation. Regulating for observability faces trade-offs between inclusiveness and depth of access, as well as line-drawing problems around the publicness of user content. And in regulating with observability, tensions arise between observability's direct role in law enforcement and its more indirect role in knowledge production and public discourse. I argue for a loose coupling between observability and regulation, and against the tendency to reduce data access to mere compliance monitoring.

1. Introduction

Social media governance has taken a regulatory turn. The policies and standards set by domi-nant platforms have become deeply politicised and are increasingly targets for government regulation. Whereas earlier social media regulation focused on the comparatively modest task of combating unlawful content, new measures such as the EU Digital Services Act (DSA) reflect far more comprehensive attempts to govern how social media moderate and curate content and align them with public interest principles.

In these reforms, the principle of transparency occupies an awkward position. On the one hand, information asymmetry is at the very foundation of platforms' economic and societal power, and disclosure regulation has therefore become a central feature of most reform programmes. In particular, much attention has been paid to the problem of algorithmic transparency, and how these hypercomplex systems can be rendered comprehensible. And yet, on the other hand, the transparency ideal itself has been undergoing a reappraisal. Since its turn-of-the-millennium heyday, a slew of failed regulatory experiments and a growing body of critical research have highlighted transparency's many limitations and failure modes; rarely the self-executing policy panacea that was expected, and often a distraction from more robust behavioural regulation.[1] In parallel, a growing body of work in critical algorithm studies has questioned whether the algorithmic transparency ideal of 'opening the black box' is at all meaningful or feasible.[2]

Synthesising these critiques, Bernhard Rieder and Jeannette Hofmann have proposed 'observability', as a pragmatic alternative to algorithm-centric models of platform transparency.[3] Their account seeks to recast platform decision making, not as a mechanistic product of monolithic algorithms but rather as the contingent outcome of complex and distributed sociotechnical systems. It decentres the explaining of

1 David Pozen, 'Transparency's Ideological Drift' (2018) 126 *Yale Law Journal* 100. Gregory Michener, 'Gauging the Impact of Transparency Policies' (2019) 79 *Public Administration Review* 136. Sandrine Baume and Yannis Papadopoulos, 'Transparency: from Bentham's inventory of virtuous effects to contemporary evidence-based scepticism' (2018) 21 *Critical Review of International Social and Political Philosophy* 169.

2 e.g. Mike Ananny and Kate Crawford, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2018) 20 *New Media & Society* 973. Nick Seaver, 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems' (2017) 4(2) *Big Data & Society.* Mike Ananny, 'Toward an ethics of algorithms: Convening, observation, probability, and timeliness' (2017) 41 *Science, Technology, & Human Values* 93.

3 Bernhard Rieder and Jeanette Hofmann, 'Towards Platform Observability'(2020) 9(4) *Internet Policy Review* <https://doi.org/10.14763/2020.4.1535> accessed 19 September 2022.

algorithms in favour of real-time, automated data access on platform behaviour and outcomes. Theirs is a pragmatic programme, which expressly ties transparency to regulation, not as an alternative but as a companion.

This paper uses the concept of observability to critique the regulation of transparency in social media law, as a specific subdomain of platform governance. In particular, it focuses on social media recommender systems, as influential algorithmic systems which have recently become the target of many new transparency requirements. My aim in doing so is to unpack the relationship between observability and regulation, taking as a starting point Rieder and Hofmann's two-way connection regulating *for* and *with* observability. How can lawmaking contribute to social media observability, and what are the challenges in doing so? What is observability's relationship to other forms of transparency? And what does it mean for observability to act as companion to regulation?

To answer these questions, I start with by introducing the domain of social media governance, and then, in detail, the concept of observability as put forward by Rieder and Hofmann. I then discuss how observability is governed in the context of social media recommender systems, and how these practices are set to be regulated in the DSA. Finally, I discuss how these observability policies might contribute to regulation.

2. Background

This section introduces the core concepts at issue in this paper: social media regulation, transparency, and observability. I start with an overview of social media regulation, and, how in instruments such as the DSA it has gradually expanded from content moderation into content curation and the regulation of recommender systems.

I then discuss the principle of transparency as a mainstay of modern governance in general, and algorithm governance in particular, which I then contrast with the critical ideal of 'observability' as developed by Bernhard Rieder and Jeanette Hofmann.

2.1 Social media regulation

Over the past decade, European governments have taken various steps to regulate social media platforms. Some policies are directed at online platforms or services in a more general sense (e.g. data protection policy, competition policy, consumer protection).[4] Others are targeted at social media and user-generated content

4 Important examples include the General Data Protection Regulation, the Digital Markets Act, and the Platform-to-Business Regulation.

regulation in particular.[5] These sector-specific rules initially developed with a strong focus on content moderation, and more specifically on the removal of unlawful content such as copyright infringements or hate speech.[6] But more recently, with proposals such as the DSA, regulatory ambitions have expanded in at least two directions: within content moderation, there is growing attention for the treatment of *lawful* content and the application of non-removal sanctions like visibility reductions.[7] Secondly, there is growing attention for the *systemic* role of recommender systems as a means of content curation; that is, the process through which platforms define and select relevant content for users.[8] Both of these developments entail a growing attention for recommender systems as a crucial point of control in social media governance.

Content moderation is the process through which platforms enforce rules applicable to user-generated content.[9] This process is carried out by (combinations of) human moderators and automated machine-learning classifiers, working apart or in concert.[10] Platforms moderate to enforce principles of national law, such as those on copyright or defamation, but also to enforce their contractual house rules, such as prohibitions on spam and nudity. Whereas unlawful content must typically be removed, lawful content can also be disciplined through alternative sanctions. Most commentators agree that some level of discretionary content moderation is socially necessary, but also that these systems are prone to error and excess.[11] At the individual level, therefore, content moderation poses threats to users' rights to information, free expression and due process.[12] And at a systemic level, content moderation may be

5 Important examples include the revised Audio-Visual Media Services Regulation, the Terrorist Content Online Regulation, the recently-proposed Political Advertising Regulation, and the Copyright Directive. In co-regulation, additionally, the Code of Practice on Disinformation and the Code of Conduct on Hate Speech.

6 Eric Goldman, 'Content Moderation Remedies' (2021) 28 *Michigan Technology Law Review* 1. João Quintais, 'The new Copyright in the Digital Single Market Directive: a critical look' (2020) 42 *European Intellectual Property Review* 28.

7 Chapter 5 above.

8 Kerstin Thorson and Chris Wells, 'Curated flows: A framework for mapping media exposure in the digital age' (2016) 26 *Communication Theory* 309.

9 Sarah Roberts, *Behind the Screen: Content moderation in the shadows of social media* (Yale University Press 2021).

10 Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) *Big Data & Society* 1 <https://doi.org/10.1177/2053951719897945> accessed 19 September 2022.

11 Tarleton Gillespie, *Custodians of the internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018).

12 Christina Angelopoulos and others, 'Study of fundamental rights limitations for online enforcement through self-regulation' (Research Report Institute for Information Law 2015). Retrieved from < pure.uva.nl/ws/files/8763808/IVIR_Study_Online_enforcement_through_self_regulation.pdf> accessed 16 September 2022.

biased so as to systemically oversanction or undersanction certain types of content, and impose disparate impacts especially on marginalised groups.[13] In recent years, platforms have developed new strategies focused on visibility management and reduction, which leave the moderated content in place but reduce its prominence in recommender systems.[14] These visibility reduction strategies have become central to the governance of controversial content such as disinformation and political extremism, but they have also been criticised for being especially opaque and fuelling anxieties about untraceable 'shadow banning'.[15]

Content curation is the process through which platforms define and select relevant items for their users, and 'filter abundance into a collection of manageable size'.[16] This is achieved through automated recommender systems.[17] These systems typically use machine-learning techniques to optimise for user engagement, though in practice their design involves a complex weighing of different competing interests and exigencies. Platforms enjoy substantial leeway in how they afford and measure engagement, and can tweak and redesign these techniques over time in order to steer the service towards desired outcomes—from commercial considerations such as user satisfaction, retention and conversion to political, reputational and regulatory considerations like mitigating disinformation or hate speech.[18]

In recent years this process of curation has become politicised, attracting scrutiny from governments and civil society as well as from the content creators who depend on recommender visibility for their livelihood.[19] One important concern is algorithmic bias and discrimination; recommender systems may reflect and entrench existing societal inequalities.[20] In platform advertising, for instance, researchers have demonstrated that Facebook ad targeting for housing ads systematically disadvantages

13 Rachel Griffin, 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality', SSRN Draft Paper (2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064738> accessed 19 September 2022. Rachel Griffin, 'The Sanitised Platform' (2022) 13 *JIPITEC* 36.

14 Tarleton Gillespie, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 *Social Media+ Society* <https://doi.org/10.1177/205630512211175> accessed 19 September 2022.

15 Chapter 5 above.

16 Thorson and Wells, 'Curated flows' (n 8).

17 Thomas Poell, David Nieborg and Brooke Erin Duffy, *Platforms and cultural production* (John Wiley & Sons 2021).

18 Tarleton Gillespie, 'The relevance of algorithms', in Tarleton Gillespie and others (eds), *Media Technologies: Essays on Communication, Materiality, and Society* (The MIT Press 2014). Poell, Nieborg and Duffy,' Platforms and cultural production' (n 17).

19 Chapter 2 above (Leerssen, 'The Soap Box as a Black Box'). Natali Helberger, 'The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power' (2022) 8 *Digital Journalism* 842.

20 Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press 2018).

ethnic minorities.[21] Another common concern is commercial discrimination, such as when platforms self-preference their own products to gain an unfair advantage over rivals in complementor markets.[22] Specific to social media, recommender systems also raise media policy concerns, concerning e.g. media diversity, child protection, and the availability of high quality and public interest content such as news and educational material.[23] Conversely, content curation is also accused of systematically 'amplifying' low-quality content and extreme, polarising content.[24] At a systemic level, content curation is accused of lowering diversity in individual media diets (associated with the theory of 'filter bubbles') and discouraging cross-group interaction and exchange (associated with the theory of 'echo chambers').[25] Again, however, as with moderation, the push for greater platform responsibility in curation has also provoked counterreactions against possible overreach. Natali Helberger warns that the push by governments to hold platforms responsible for curation outcomes risks further entrenching their systemic opinion power, and in this sense could *harm* media pluralism more than it helps.[26] Together with Jo Pierson and Thomas Poell she argues for a more hands-off approach in which governments aim to set public values for content curation not through conventional command-and-control but in close cooperation with users and civil society.[27]

The DSA is the first major legislation to tackle these issues. For content moderation, it introduces a comprehensive due process framework covering all moderation actions, including actions against lawful content and through non-removal remedies. In short, this framework requires platforms to clearly define their own content moderation rules, to enforce them consistently, to notify and explain each moderation decision to those affected, and to arrange internal and external appeals procedures for these decisions.[28] For content curation, its rules are far less precise. They require large

21 Muhammad Ali and others, 'Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes' (2019) 3 *Proceedings of the ACM on human-computer interaction* 1.

22 Ariel Ezrachi and Maurice Stucke, 'Virtual competition' (2016) 7 *Journal of European Competition Law & Practice* 585.

23 Natali Helberger, Katharina Kleinen-Von Königslöw and Rob van der Noll , 'Regulating the new information intermediaries as gatekeepers of information diversity' (2015) 17 *info* 50.

24 Daphne Keller, 'Amplification and its discontents: why regulating the reach of online content is hard', Knight First Amendment Institute (8 June 2021) < https://knightcolumbia.org/content/amplification-and-its-discontents> accessed 25 September 2022.

25 Eli Pariser, *The Filter Bubble: What the internet is hiding from you* (Penguin 2011). Axel Bruns, *Are Filter Bubbles Real?* (Polity Press 2019). Frederik Zuiderveen Borgesius and others, 'Should we worry about filter bubbles?' (2016) 5(1) *Internet Policy Review* <https://doi.org/10.14763/2016.1.401> accessed 20 September 2020.

26 Helberger, 'The political power of platforms' (n 19).

27 Natali Helberger, Jo Pierson and Thomas Poell, 'Governing online platforms: From contested to cooperative responsibility' (2018) 34 *The Information Society* 1.

28 DSA, Articles 14, 17, 20 and 21.

platforms—defined as those with more than 45 million monthly average active EU users—to regularly assess and mitigate so-called 'systemic risks', an extremely broad category which covers the dissemination of illegal content, the protection of fundamental rights (with specific reference to media freedom and media pluralism), the protection of civic discourse and electoral processes, and the protection of minors.[29] Recommender systems in particular are mentioned as a possible source of risk, and a solution space for risk mitigation.[30] Platforms will have to publish yearly reports on their diagnosis and mitigation of risks, and the European Commission may issue guidance as to how these duties are carried out.[31] In this way, the DSA sets the stage of an open-ended co-regulatory process of standard setting on content curation. All these rules are accompanied by a slew of transparency requirements, which I return to in Section 3 below.

All these trends combine to make recommender systems a new battleground in social media regulation. As a site of curation they define, in a very literal sense, what it means to be relevant online. As sites of moderation, they police the boundaries of acceptable expression and condemn perceived transgressors to obscurity. With the DSA, the law has only started in the most general terms to articulate public interest principles for these systems, and in doing so it raises difficult questions about the appropriate balance between individual user choice and public ordering, and the role of the state and the law in online media governance. For this reform effort, an overarching problem is that recommender systems are deeply opaque. Much regulatory and scholarly effort has therefore been trained at 'opening the black box' and uncovering the inner workings of these algorithmic systems.[32] Below I discuss the regulatory politics of algorithmic transparency and then Rieder and Hofmann's critical alternative: observability.

2.2 From transparency to observability

Transparency has been described as a 'quasi-religious principle' of modern governance.[33] By providing external access to internal information, transparency promises to make organisations more accountable and efficient.[34] Although this

29 DSA, Articles 33, 34 and 35.

30 DSA, Articles 34(2)(a) and 35(1)(d).

31 DSA, Article 35(3).

32 Frank Pasquale, *The Black Box Society: The secret algorithms that control money and information* (Harvard University Press 2015). Karen Yeung, 'Algorithmic regulation: A critical interrogation' (2018) 12 *Regulation & Governance* 505.

33 David Heald and Christopher Hood and David Heald (eds.). In *Transparency: The key to better governance?* (Oxford University Press for the British Academy 2006).

34 Mikkel Flyverbom, *The Digital Prism: Transparency and managed visibilities in a datafied world* (Cambridge University Press 2019).

principle of transparency can be traced as far back as the progressive era or even the enlightenment (then referred to as 'publicity'), its popularity reached new heights around the turn of the millennium.[35] Since then, a reappraisal has started to take place, questioning the accuracy of transparency's products, the effects they produce, and the costs they impose.[36] Perhaps the most common criticism is that transparency too often serves as an *alternative* to substantive or behavioural regulation, or even as a wedge against it.[37] In these instances transparency is invoked as a reason not to impose binding duties, on the often unrealistic assumption that 'more information will lead to better behavior'—reflecting a neoliberal faith in market competition and individual choice as superior ordering mechanisms.[38] A more mature engagement with transparency, David Pozen argues, should 'desacralise' this article of faith, and approach it not as a miracle cure but rather as a support or catalyst for regulation.[39]

At the same time, transparency has become central to many analyses of digital policy and platform regulation. Platforms are in essence information services, whose very business model and regulatory power consist in the appropriation and exploitation of large datasets.[40] Due to the profound information asymmetries these services create, many regulatory proposals aim to adjust this imbalance and open up access to outsiders.[41] Transparency, then, is often seen as an indispensable precondition for the exercise of democratic control over these services.[42] Transparency has gained further relevance due to the rise of machine-learning algorithms in platform governance, in such areas as content moderation and curation.[43] An emerging field of algorithm governance treats opacity as one of the distinctive features of machine learning; one

35 Emmanuel Alloa, 'Transparency: A magic concept of modernity' in Emmanuel Alloa and Dieter Thomä (eds.), *Transparency, society and subjectivity* (Palgrave Macmillan 2018). Pozen, 'Transparency's ideological drift' (n 1). Robert Gorwa and Timothy Garton Ash, 'Democratic Transparency in the Platform Society' in: Nate Persily and Joshua Tucker (eds.), Social Media and Democracy: The State of the Field and Prospects for Reform (Cambridge University Press 2020). But c.f.: Emmanuel Alloa, 'Why transparency has little (if anything) to do with the age of enlightenment' in Emmanuel Alloa (ed.), *This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor* (Leuven University Press 2022).

36 e.g. Pozen, 'Transparency's Ideological Drift' (n 1). Michener, 'Gauging the impact of transparency policies' (n 1). Baume and Panadopoulos, 'Transparency: from Bentham's inventory of virtuous effects to contemporary evidence-based scepticism' (n 1).

37 Ibid. See also: Catharina Lindstedt and Daniel Naurin, 'Transparency is Not Enough: Making Transparency Effective in Reducing Corruption' (2010) 31 *International Political Science Review* 301.

38 Flyverbom, *The Digital Prism* (n 34).

39 Pozen, 'Transparency's ideological drift' (n 1).

40 Julie Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press 2019).

41 Gorwa and Garton Ash, 'Democratic transparency in the platform society' (n 35).

42 Ibid.

43 Gorwa, Binns and Katzenbach, 'Algorithmic content moderation' (n 35).

of its main projects, hence, is to 'open the black box' and subject machine learning to outside scrutiny.[44] Since these algorithms are generally too complex for human-scale reasoning, exhaustive explanations are not feasible and instead the challenge is to render accounts which highlight the most salient factors— salience being, of course, in the eye of the beholder.[45] All this leads to uneasy ambivalence around transparency in platform governance; an ideal past its prime, but in some sense more relevant than ever. The crisis of faith meets a counter-reformation.

To capture some of the distinct challenges and opportunities for transparency in platform governance, Bernhard Rieder and Jeanette Hofmann have proposed the term 'observability'.[46] Under this banner, they make several recommendations for disclosure regulation which responds to the particular epistemic challenges posed by platforms and their automated decision-making systems. Below I will first describe observability's theoretical and conceptual underpinnings, and then its regulatory principles, which I will subsequently apply to the DSA.

Rieder and Hofmann's approach is informed by two strains of scholarship. First, they build on the critical transparency studies cited above, and its problematisation of transparency as an objective source of truth. In addition, Rieder and Hofmann also draw on Science and Technology Studies (STS) and critical algorithm studies, which have problematised the ideal of algorithmic transparency.[47] Recent work in this field has sought to resist the preoccupation with algorithms as sites of power and objects of study, and, with it, the dominant frame of 'opening the black box'. STS instead approaches platforms' automated decision-making as a sociotechnical process, where meanings and outcomes are shaped not only by algorithmic design, but in large part also by the actions of users and other stakeholders who interact with these systems.[48] 'As use practices change, algorithmic decision models change as well', Rieder and Hofmann note, and platforms

44 Pasquale, *The Black Box Society* (n 32). Yeung, 'Algorithmic regulation' (n 32). Danielle Citron and Frank Pasquale, 'The scored society: Due process for automated predictions' (2014) 89 *Washington Law Review* 1.

45 Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2017) 3(1) Big Data & Society <https://doi.org/10.1177/2053951715622512> accessed 15 September 2022. Lilian Edwards and Michael Veale, 'Slave to the algorithm? Why a "Right to an Explanation" is probably not the remedy you are looking for' (2017) 16 *Duke Law & Technology Review* 18. Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31 *Harvard Journal of Law and Technology* 841

46 Rieder and Hofmann, 'Toward platform observability' (n 3).

47 e.g. Ananny and Crawford, 'Seeing without knowing' (n 2). Ananny, 'Towards an ethics of algorithms' (n 2). Seaver, 'Algorithms as Culture' (n 2). Gillespie, 'The relevance of algorithms' (n 18).

48 Rieder and Hoffmann, 'Toward platform observability' (n 3).

neither fully understand nor fully control these complex sociotechnical processes.[49] Explaining the algorithm isn't just technically challenging; it is insufficient as a window onto platforms' automated decision-making practices. Instead of just opening the black box, understanding platform behaviour requires attention to their social embeddedness; how their technologies interact with social contexts and communities of practice, and how these forces shape one another over time.

How, then, does observability start to address these shortcomings? Starting with its semantics, Rieder and Hofmann emphasise above all that observability is a pragmatic concept; whereas the transparency metaphor refers to a physical property, ascribed to materials, observability contains the potential for an *act*, carried out by observers. If transparency describes a view from nowhere, observability draws attention to the viewer(s) and their perspective(s). If transparency pretends to objectivity, therefore, observability highlights subjectivity and the communicative work of disclosure; it intends to 'draw attention to and problematise the process dimension inherent to transparency as a regulatory tool.'[50] Transparency is passive, static, and pretends to objectivity; observability is active, pragmatic, and departs from subjectivity.

I will note another semantic difference, which Rieder and Hofmann do not belabour explicitly but which still seems to inform their analysis. Namely: transparency and observability suggest different objects or directionalities. The transparency metaphor implies an act of seeing *through* materials or boundaries, and hence *inside*

49 Rieder and Hofmann, 'Toward platform observability' (n 3) 8.

50 Rieder and Hofmann also describe observability as foregrounding the *mediated* nature of disclosure practices. Yet it strikes me that their chosen metaphor of 'observability' does not necessarily work in their favour here. After all, observability draws on the same language of sight, the same coupling of seeing and knowing, as does transparency. If anything, transparency suggests the *more* mediated perspective of the two metaphors, since it describes a view that is literally mediated by a diaphanous material or medium. Hence, the language of transparency *can* in fact accommodate critical interrogations of mediation, for instance in Flyverbom's "digital prism" of distorted and refracted sight, or Emmanuel Alloa's biblical invocation of "seeing as through a glass, darkly". For Ida Koivisto, the transparency metaphor is not merely iconoclastic but icono-ambivalent; it necessarily implies a mediation, even though the goal of this mediation is to render its target as clearly as possible and hence to escape notice. Observability, by contrast, suggests no mediation at all—at least in its semantics—beyond, perhaps, in its aforementioned pragmatism, with its attention to perspective and interpretation over objective access to reality. See: Emmanuel Alloa (ed.), *This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor* (Leuven University Press 2022). Ida Koivisto, *The Transparency Paradox: questioning an ideal* (Oxford University Press 2022). Flyverbom, *The Digital Prism*. Ananny and Crawford, 'Seeing without knowing' (n 2).

the organisation or system under scrutiny.[51] This is why the metaphor of 'opening the black box' is intrinsically connected to that of transparency; they both seek to look inside and make externally visible that which is internal. Observability, by contrast, lacks this directionality; it does not imply seeing *through*, but merely seeing. It is far more capacious. Transparency always peers inside but observability can also look at, on, under, around or across its object.[52] Applied to algorithmic systems, therefore, observability suggests we expand our view from the algorithm as such to encompass other aspects, such as inputs, outputs, and interventions. If transparency would have our nose pressed up against the glass, then observability permits us to take in the surroundings.

	Transparency	**Observability**
Denotes:	Material property (diaphaneity) Capacity to see *through* and *inside*	Practice (observing) Capacity to see
Connotes:	Capacity to inspect Objective, passive, static Accounting of (internal) reasons Algorithms (technical perspective)	Capacity to regard, locate Subjective, active, pragmatic Awareness of outcomes, effects, context Assemblages (sociotechnical perspective)

Table 4: Semantics of transparency and observability

In sum, then, these two characteristics of observability—its pragmatism, and its decentered directionality—respond to two major critiques of algorithmic transparency reforms. Its pragmatism dispels the naïveté of transparency as a source of objective (empirical) truth, and calls attention to the selectivity, subjective reception and contingent effects of disclosure. Its decentered directionality averts the focus on algorithms as objects of study, and encourages a more holistic and contextual assessment of algorithmic decision-making as embedded in a social context. I now turn to the practical, policy-oriented implications of this reformulation.

2.3 Observability as a regulatory program

To achieve observability, Rieder and Hofmann outline three main principles. The first is to 'expand the normative horizon' as to what transparency can achieve.[53] Platformisation represents a fundamental reshaping of many societal domains; its information asymmetries challenge not only regulatory enforcement, fair dealing,

51 A detailed commentary on the directionality of transparency has been written by David Heald. He uses transparency more capaciously, with an inward but also an outward meaning. This outward variant is rarely invoked in the contexts of algorithmic and platform governance, and I therefore restrict myself to the more salient inward variant. See: David Heald, 'Varieties of transparency' in David Heald and Christopher Hood and David Heald (eds.). In *Transparency: The key to better governance?* (Oxford University Press for the British Academy 2006).

52 Ananny, 'Toward an ethics of algorithms' (n 2).

53 Rieder and Hofmann, 'Toward platform observability' (n 3) 10.

or user choice, but entail a more profound shift in society's capacity for knowledge production. Even platforms themselves do not possess total knowledge as to their services' functioning, but they do possess the data which is necessary to study them. In this way, platformisation 'deprives society of a crucial resource for producing knowledge about itself'.[54] More concretely, this suggests an approach to transparency which goes beyond regulatory auditing or individual disclosures, and also aims to empower researchers and civil society actors to gain access to platforms resources. And in terms of substance, it reimagines transparency regulation not as merely divulging knowledge, but as forcing access to the means of knowledge production.

The second principle is to 'observe platform behaviour over time'.[55] Given the volatility of platform ecosystems—their constant adaptation to changing social contexts—disclosures need to move beyond the conventional 'snapshot logic' of periodical auditing or reporting.[56] Instead, society requires structures to study platforms over time, and ideally in real-time. Rieder and Hofmann list four access methods which reflect this approach: (1) Data access agreements, where specific researchers are granted access to select datasets, typically under conditions of confidentiality; (2) Accountability Interfaces, which provide automatic transparency functions to a wider (public) audience; (3) Developer APIs, which are similar to accountability interfaces but are designed primarily with commercial usage in mind rather than accountability per se; and (4) data scraping, where researchers collect platform data via end user interfaces and independently of the platform. The following section will discuss examples of each in greater detail, in the context of social media recommender systems.

The third principle is to 'strengthen foundations for collaborative knowledge creation'. Part of observability's pragmatic approach is an attentiveness to the different information needs of specific actors in making sense of platform data. Transparency rules which prefigure the relevant facts or norms under scrutiny will likely fail to accommodate these different perspectives, needs and interests. Access to the underlying platform data, through interfaces such as the above, can help 'different actors to develop their own observation capacities, adapting their analytical methods to the questions they want to ask'.[57] Still, Rieder and Hofmann recognise that meeting third party needs is a key challenge for observability policy; it 'raises the complicated question of how data and analytical capacities should

54 Rieder and Hofmann, 'Toward platform observability' (n 3) 11
55 Rieder and Hofmann, 'Toward platform observability' (n 3) 7.
56 Rieder and Hofmann, 'Toward platform observability' (n 3) 13.
57 Rieder and Hofmann, 'Toward platform observability' (n 3) 20.

be made available, to whom, and for what purpose'.[58] Beyond mere *access* to data, effective observability also calls for institutional capacity-building and adequate funding of researchers and other societal watchdogs.

Rieder and Hofmann present observability as tightly coupled to regulation. First, regulation is necessary to achieve observability policies; platforms are unlikely to carry them out effectively of their own accord, at least not without the threat of government regulation to motivate them. Second, observability should be conceived as a companion to regulation. Here, Rieder and Hofmann are responding to criticisms that transparency often functions as a deregulatory wedge against substantive rulemaking. Far from an alternative to regulation, they emphasise that the ultimate goal of observability should be 'to assess platform behaviour against public interest norms' and to 'undergird the regulatory response to the challenges platforms pose'.[59]

The following sections will trace this two-way relationship between regulation and observability, focusing on the specific case of social media: First, how can social media regulation contribute to observability? Second, how can observability contribute to social media regulation? As we will see, regulating for observability is by no means theoretical but in fact an important feature of the new DSA.

3. Regulating social media for observability

The following section applies the principle of observability to an especially controversial category of platforms' automated decision-making: social media recommender systems. What does observability for these tools look like in practice? I first discuss the main observability tools described by Rieder and Hofmann, and how these apply to social media recommender systems. I then review the DSA's key provisions on transparency. As we will, several of these resonate with the principle of observability, whilst also surfacing important challenges for this regulatory project.

3.1 Observability in social media governance

Rieder and Hofmann describe a number observability resources as possible targets for regulation. What these have in common is that they permit the observing of platform behaviour over time, or indeed in real-time, allowing their disclosures to keep pace with the volatility and dynamism of platform ecosystems—unlike the 'snap-shot logic' of periodical auditing or reporting. Possible approaches include: (1) Data access

58 Rieder and Hofmann, 'Toward platform observability' (n 3) 21.
59 Rieder and Hofmann, 'Toward platform observability' (n 3) 23.

agreements, (2) accountability interfaces, (3) developer APIs; and (4) data scraping. Below I discuss how these categories apply to social media and their recommender systems.

Data access agreements are confidential arrangements, where specific researchers are granted access to select datasets. Their usage of this data can be restricted by legal means (i.e. non-disclosure agreements, or 'NDAs') and technical means (secure operating environments, differential privacy techniques). Data sharing has a long history in social media, starting with *ad hoc* arrangements but gradually being formalised through programs such as Facebook's Social Science One. These confidential arrangements can enable researchers to work with privacy-sensitive data. For social media, this can include detailed audience analytics on individual media diets, in order to study how users discover content in practice and interact with recommender systems over time. Confidential arrangements can also enable more detailed engagement with recommender algorithms, for instance through 'code audits', which allow experts to examine the internal source code of automated systems, as well as 'sandboxing' which permits them to experiment, in a restricted or simulated environment, with different scenarios and configurations.[60]

Accountability interfaces are automated solutions which disclose non-sensitive data to a wider (public) audience. One important example from social media are ad archives: public repositories documenting advertisements sold on the service. Although the self-regulatory versions have been criticised for various inaccuracies and omissions, these tools have nonetheless found uptake amongst journalists and academics, especially in countries where online political advertising is prevalent such as in the UK and US.[61] Another important accountability interface for social media recommender systems is Facebook's CrowdTangle, which provides high-level aggregate engagement statistics not for advertisements but for organic content.[62] To mitigate privacy concerns, it does not cover all Facebook content and is instead limited to content from public pages

60 Christian Sandvig and others, 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms' (2014), Paper presented to Data and Discrimination: Converting Critical Concerns into Productive Inquiry <https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf> accessed 26 September 2022. Brent Mittelstadt, 'Automation, algorithms, and politics: auditing for transparency in content personalization systems' (2016) 10 *International Journal of Communication* 12.

61 Chapter 4 above (Paddy Leerssen, Tom Dobber, Natali Helberger and Claes de Vreese, 'News from the ad archive: how journalists use the Facebook Ad Library to hold online advertising accountable' (2021) *Information, Communication and Society* <https://doi.org/10.1080/1369118X.2021.2009002>). Michael Bossetta, 'Scandalous design: How social media platforms' responses to scandal impacts campaigns and elections' (2020) 6(2) *Social Media+ Society* <https://doi.org/10.1177/2056305120924777> accessed 16 September 2022.

62 Richard Rogers, 'Social media research after the fake news debacle' (2018) 11 *Partecipazione e conflitto* 557.

and from an (undefined) selection of 'influential' users.[63] It is worth noting that some accountability interfaces are not strictly public but also limit participation in some way, for instance by requiring a proof of identity, affiliation with a journalistic enterprise or consent to an NDA; a sliding scale, then, between fully public interfaces and once-off data grants.[64]

Developer APIs are similar to accountability interfaces, but differ in their purpose: they are not designed for purposes of transparency or research, but rather for commercial actors. Still, researchers can glean useful data from them. Developer APIs have been heavily restricted over the past years, in a development known as the 'APICalypse'.[65] For instance, Bernhard Rieder, Oscar Coromina and Ariadne Matamoros-Fernandez were previously able to use YouTube's Web-API to survey many millions of videos and create a general overview of the most popular categories of content on the service, but note that such research is now 'difficult to replicate' because YouTube 'no longer seems to issue similarly generous [access tokens] for new research projects'.[66] Interestingly, over the past year social media platforms have created new APIs specifically for researchers, such as Twitter's Academic Research program, blurring the lines between accountability and developer interfaces.[67] Therefore, following Van der Vlist et al, it may instead be more useful to speak in general terms of an emerging 'API governance' which manages both commercial and civil society access.[68]

Data scrapers collect information directly from the platform's end-user interfaces, typically through automated bot accounts ('sock puppets') or with the help of volunteers ('data donations'). Whereas the previous categories all occur at the platforms' leisure, scraping is more adversarial and can proceed without the platform's knowledge or assent. Scraping research has been an important method to study recommender systems,

63 Chris Miles, 'What data is Crowdtangle tracking?' (2022) *CrowdTangle* <https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking> accessed 26 September 2022. [https://perma.cc/3HXR-FNPG].

64 For Poell, Nieborg and Duffy, this combination of technical and contractual parameters is characteristic of platforms' API-based regulation, including new hybrid forms such as 'badges', 'tokens' or 'certications'. Poell, Nieborg, Duffy, *Platforms and Cultural Production* (n 17). See also: Anne Helmond and Fernando van der Vlist, 'Social media and platform historiography: Challenges and opportunities' (2019) 22 *TMG–Journal for Media History* 6.

65 Axel Bruns, 'After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research' (2019) 22 *Information, Communication & Society* 1544.

66 Bernhard Rieder, Òscar Coromina and Ariadna Matamoros-Fernández, 'Mapping YouTube: A quantitative exploration of a platformed media system' (2020) 25 *First Monday* 8.

67 Adam Tornes, 'Enabling the future of academic research with the Twitter API', *Twitter Developer Blog* (21 January 2021) <https://developer.twitter.com/en/blog/product-news/2021/enabling-the-future-of-academic-research-with-the-twitter-api> accessed 26 September 2022.

68 Fernando van der Vlist and others, 'API Governance: The Case of Facebook's Evolution' (2020) 8(2) *Social Media+ Society* <https://doi.org/10.1177/2056305121221086228> accessed 20 September 2022.

being used in research on political advertising, media diversity, algorithmic bias and discrimination, and shadow banning.[69] Scraping has become highly controversial, since it can also enable abuse, consisting mainly in the the illicit collection of sensitive personal. One clearly unlawful example is ClearviewAI's mass scraping of social media for facial recognition purposes.[70] But civil society usage can pose its own legal grey areas and ethical quandaries.[71] Platforms are rarely interested in nuance, however, and prohibit scraping as a rule in their Terms of Service regardless of possible public interest defences.[72] Academics have mounted litigation in several courts to challenge these restrictions, primarily in the US, but these remain largely inconclusive.[73]

In sum, what can these observabilities reveal about social media? Most tell us little about their recommender *algorithms*, but much about recommender *systems* as these operate in practice. They rarely provide conclusive answers on algorithmic reasoning or causality, but they do tell us about their outcomes. They allow researchers to survey online media ecologies; across personally curated flows, social media observability permits us to see what others are seeing.

One way to understand social media observability is that it emulates certain basic affordances of the non-personalised mass media; maintaining a stable record of significant public communications (such as through data scraping and ad archiving), documenting important alterations or interventions to this record (such as visibility restrictions and other moderation actions), and locating these communications within specific contexts and audiences (such as through data scraping or audience analytics). Yet, we have seen that observing needs regulating; most voluntary arrangements are incomplete and overly dependent on the platforms they scrutinise.

69 Sandvig and others, 'Auditing algorithms' (n 60). Brent Mittelstadt, 'Automation, algorithms, and politics: auditing for transparency in content personalization systems' (2016) 10 *International Journal of Communication* 12. Balazs Bodó and others, 'Tackling the Algorithmic Control Crisis: The Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents' (2018) 19 *Yale Journal of Law and Technology* 133. Eduardo Hargreaves and others, 'Fairness in Online Social Network Timelines: Measurements, Models and Mechanism Design', 127 *Performance Evaluation Review* 15. Márcio Silva and others, 'Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook' (2020) *WWW '20: Proceedings of the Web Conference 2020* 224.

70 Isadora Rezende, 'Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective' (2020) 11 *New Journal of European Criminal Law* 375.

71 Gabriel Fair and Ryan Wesslen, 'Shouting into the void: A database of the alternative social media platform Gab' (2019) 13 *Proceedings of the International AAAI Conference on Web and Social Media* 608. Jacquellena Carrero, 'Access Granted: A First Amendment Theory of Reform of the CFAA Access Provisions' (2020) 120 *Columbia Law Review* 131.

72 Cohen, *Between Truth and Power* (n 40).

73 Benjamin Sobel, 'A New Common Law of Web Scraping' (2021) 25 *Lewis & Clark Law Review* 147.

3.2 Observability in the Digital Services Act

The DSA is the first major legislation to regulate platform observability. Whereas earlier transparency rulemaking has focused on conventional public reporting methods, the DSA contains provisions to regulate real-time data access specifically for researchers and other civil society actors.[74] Its most important provision to this effect is Article 40 on research access, which I explain further below. In addition, the DSA also contains specific observability rules on ad archives and on content moderation archives.[75] The discussion below focuses on Article 40 because it is a more general framework, not aimed at any specific content but instead laying down a catch-all procedure to request data access for researchers.

Before proceeding it should also be noted that the DSA contains other types of transparency rules besides, which are not expressly designed for observability. For instance, there is also an explanation rule for recommender systems, which requires platforms to disclose the 'main parameters' used to rank content—precisely the type of disembedded and algorithm-centric rulemaking which observability aims to surpass.[76] The DSA also contains more conventional rules on contractual transparency and due process, user labelling, and periodical reporting on various topics.[77] Although these rules do not rise to the level of observability, neither are they entirely irrelevant to the

74 On these conventional reporting rules, see generally: Daphne Keller and Paddy Leerssen, 'Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation', in: Persily N. and Tucker J (eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge University Press 2020).

75 Article 39 DSA would help to improve and expand ad archive practices which already take place through self-regulation and private ordering, making it a less momentous or radical break from current practice than Article 40 DSA. Article 24(5) DSA on content moderation archives does describe a novel proposal to archive the 'Statements of Reason' that platforms must now issue for each moderation action. Although moderation archiving has been a long-standing demand of independent researchers, Article 22(3)'s approach to this issue is hamstrung by a lack of content-specific data to clarify what items have actually been affected. In this sense, it falls short of true of observability and instead merely reproduces the platform's own classification decisions. Even for moderation archiving, therefore, the general framework of Article 31 DSA seems the more significant development. On moderation archiving, see generally: John Bowers, Elaine Sedenberg and Jonathan Zittrain, 'Platform Accountability Through Digital "Poison Cabinets"'. Knight First Amendment Institute (13 April 2021). <https://cyber.harvard.edu/story/2021-04/platform-accountability-through-digital-poison-cabinets> accessed 16 September 2022. MacKenzie Common, 'Fear the reaper: How content moderation rules are enforced on social media' (2020) 34 *International Review of Law, Computers & Technology* 126. David Erdos, 'Disclosure, Exposure and the "Right to be Forgotten" after Google Spain: Interrogating Google Search's webmaster, end user and Lumen notification practices' (2020) 38 *Computer Law & Security Review* 105437.

76 DSA, Article 27.

77 Martin Husovec and Irene Roche Laguna, 'Digital Services Act: A Short Primer', in: Husovec and Roche Laguna, *Principles of the Digital Services Act* (Oxford University Press forthcoming 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4153796> accessed 25 September 2022.

observability project. The high-level information they provide, though likely incomplete and simplified, can still serve as a starting point for more far-reaching independent inquiry. Along these lines, an 'ecological' perspective recognises that information disclosures may amount to more than the sum of their parts through aggregation, cross-referencing, corroboration, hypothesis formation, and so forth.[78] For instance, as I have argued elsewhere, the DSA's individual due process rights might feed into social accountability, when individual users alert their findings to broader publics.[79] Transparency measures never operate in isolation but always as part of a broader information ecosystem, which may well give rise to synergies between conventional transparency rules and more innovative and ambitious observability rules. This being said, my analysis below focuses on the latter, and therefore on Article 40 on data access and scrutiny.

Article 40 is designed as follows. It regulates access to confidential data for 'vetted researchers', and access to public data for 'researchers'.[80] In addition, it also regulates data access for the DSA's public authorities, but these provisions remain out of scope for this paper.[81] Article 40 applies only to large platforms with more than 45 million monthly average active users.

The vetting procedure is initiated when a researcher submits an *application* to the Digital Services Coordinator.[82] In this application, the researcher must demonstrate:

(a) that they are affiliated with a 'research organisation', which has been defined in previous legislation as 'to conduct scientific research or to carry out educational activities'[83]

78 Seth Kreimer, 'The freedom of information act and the ecology of transparency' (2007) 10 *University of Pennsylvania Journal of Constitutional Law* 1011. René Mahieu and Jef Ausloos, 'Harnessing the collective potential of GDPR access rights: towards an ecology of transparency', Internet Policy Review (6 July 2020) <https://policyreview.info/articles/news/harnessing-collective-potential-gdpr-access-rights-towards-ecology-transparency/1487> accessed 22 September 2022.

79 Chapter 5 above.

80 DSA, Article 40(4) and 40(12).

81 DSA, Articles 40(1), 40(2) and 40(3).

82 DSA, Articles 40(4), 40(8) and 3(n). The application must be processed by the DSC of establishment, i.e. of the Member State where the platform in question has its main establishment or legal representative. Researchers can also submit to their own national DSA, who must then forward the application to the DSA of establishment after an initial assessment of conformity. See: DSA, Article 40(9).

83 This definition is taken from the Copyright Directive, and its regulation of text and data mining exceptions. In addition to the above criteria, these research organisations must also act 'in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organisation'. Furthermore, they must be organised 'on a non-profit basis or by reinvesting all the profits in its scientific research', or 'pursuant to a public interest mission recognised by a Member State'.

(b) that they are independent of commercial interests;
(c) the funding of their research;
(d) that they are capable of fulfilling data security and confidentiality requirements corresponding to each request and to protect personal data, and to describe in their request the appropriate technical and organisational measures that they have put in place to this end; and
(e) that their request is 'necessary for, and proportionate to, the purposes of their research' and furthermore that it will contribute to the understanding of Article 34 and 35's purpose of mitigating;
(f) that the research will be carried out for the aforementioned purposes (an apparent repetition of the foregoing requirement), and, finally,
(g) that they have committed to make their research results available publicly and free of charge.[84]

Upon receipt of a valid request, the DSC must, in principle, forward it to the platform. However, they may only do if they conclude that the data will be used for the sole purpose 'to monitor compliance with' the DSA and so long as the request 'takes due account of the rights and interests' of the platform and its users, which include data protection, service security and 'the protection of confidential information'.[85] There is no formal discretion for the regulator, but these open standards do leave quite some interpretive leeway.

The DSC's request is not the final word, however. The platform may respond by requesting amendments, if they consider they are 'unable' to give access to the data requested, either because (a) 'they do not have access to the data', or (b) disclosure will 'lead to significant vulnerabilities in the security of their service or the protection of confidential information, in particular trade secrets'.[86] This request for amendments must propose one or more alternative means through which the purpose of the request may be satisfied instead.[87] The DSC decides on these amendments. Then, the platform must comply by facilitating and providing access through 'appropriate interfaces specified in the request, including online databases or application programming interfaces'.[88] This could be interpreted as a way for the law to require platforms to regulate automated observability tools, such as content archives, analytics dashboards or APIs.

84 DSA, Article 40(8).
85 DSA, Article 40(4).
86 DSA, Article 40(5).
87 DSA, Article 40(6).
88 DSA, Article 40(7).

An entirely separate access rule for 'researchers' is tucked away in its penultimate subparagraph, Article 40(12). This rule, which is known in Brussels circles as the 'CrowdTangle provision', sets a lower standard of review for access to publicly available content. It requires that large platforms 'give access without undue delay to data […] provided that the data is publicly accessible in their online interface by researchers'.[89] The accompanying recitals clarify the types of data being referred to here, 'for example on aggregated interactions with content from public pages, public groups, or public figures, including impression and engagement data such as the number of reactions, shares, comments'.[90] The 'researchers' being referred to here are a slightly broader category than the aforementioned 'vetted researchers'; they must fulfil all the same requirements listed above *except* (a) affiliation with a research organisation, and (g) publication.[91] Curiously, this provision does not specify any actual procedure for these researchers to request access or to demonstrate their eligibility.

At a minimum, it seems, Article 40(12) might be invoked as a negative duty or non-interference rule; that platforms refrain from interfering with researchers studying publicly accessible information.[92] The platform might then be expected to call on researchers to demonstrate their compliance, before imposing any restrictions or obstructions.[93] A more ambitious reading would invoke Article 40(12) as a positive duty to proactively disclose data to researchers, and accredit them for this purpose. Some might see a contradiction here: if the data in question is already publicly accessible, then what additional access would this article provide? The answer may lie in the disclosure *format* and *aggregation*. Tools like CrowdTangle document data that is often technically public already, but still difficult to oversee in a systemic fashion. Still, this approach leaves some open questions as to what qualifies as 'public' data, which I return to below.

3.3 Discussion: observability of what and for whom?

The DSA surfaces some difficult questions for the regulation for social media observability. Below I focus on two key issues: recipients and substance. In other words, observability for *whom*, and of *what*? Between its two access rules, we see two approaches to these questions: confidential and exclusive access (for 'vetted researchers') versus public and inclusive access (for 'researchers').

89 DSA, Article 40(12).

90 DSA, Recital 96.

91 DSA, Article 40(12).

92 Recital 98 supports this reading by emphasising the negative duty that large platforms 'should not prevent' researcher access. Further on, however, the same recital returns to the positive duty 'should provide access'.

93 Paddy Leerssen, 'Platform research access in Article 31 of the Digital Services Act: Sword without a shield?' (Verfassungsblog 7 September 2021) <https://verfassungsblog.de/power-dsa-dma-14/> accessed 5 November 2022.

For the confidential approach, one of the main risks is that its application procedure will be too cumbersome to serve a meaningful number of researchers. Every researcher must undergo a detailed vetting procedure, and even though the law entitles them to an answer 'within a reasonable period', it remains to be seen whether DCSs will be able to live up to this demand.[94] (Since many large social media services have their establishment in Dublin, the Irish DSC will have an especially important role to play. This a potential cause for concern, since Ireland has in other areas of digital policy been accused of overly permissive or even anti-regulatory enforcement policies.) By virtue of the statute's broad exceptions, platforms will also have several means to complicate and protract proceedings.

Realising Article 40's full potential and maximising its uptake may therefore require some degree of standardisation and, ideally, automation. One starting point may be to maintain a public register of previously-successful requests, so that new applicants can follow similar templates to access the same or similar data, without triggering *de novo* review of all the underlying merits. Over time, by establishing these types of precedents, regulators could force platforms to maintain a set of APIs and other automated resources for many researchers to access, for which compliance and eligibility criteria would be well-established and applications could be processed with relative ease. In this way, Article 40 DSA might then approach a form of API governance, regulating transparency by design.[95] Yet, this vision still depends heavily on the DSC's effort and ingenuity. Given the complexity of the subject matter, as well as the institutional constraints on funding and expertise, all this may be too much to expect from regulators, at least on the short term. For now a more piecemeal and *ad hoc* approach appears likely.

In this respect, Article 40(12) DSA is perhaps the more flexible tool, since it relies less on the DSC of establishment to become useful; it is not only a sword for regulators to issue demands, but also a shield for researchers to ward off interference. Still, compared to the confidential access framework, this provision has unfortunately received relatively short thrift. Sparse on details and convoluted in its phrasing, it leaves many important questions open which hinder its application in practice. One important question it leaves open, as mentioned, is how researchers are to request data or become accredited; no procedure is specified. On paper, it does address a substantially broader audience than Article 40's confidential framework; by dropping the requirement that researchers be affiliated with a 'research organisation', it opens it up to a broader set of actors, including journalists, activists, and political campaigners.

94 DSA, Article 40(3).

95 Fernando van der Vlist and others, 'API Governance: The Case of Facebook's Evolution' (2020) 8(2) *Social Media+ Society* <https://doi.org/10.1177/2056305121221086228> accessed 20 September 2022.

Still, the substantive eligibility requirements for researchers, such as demonstrating subsidiarity and proportionality, could be cumbersome in practice (compared to, for instance, strictly public tools such as ad archives), and leave no room for derogation. Unfortunately the DSA provides no statutory basis to regulate fully public resources, limiting its overall flexibility.

The question of *who* gets access is of course interconnected with the question of *what* data they get access to. For Article 40's confidential framework, access is unrestricted in principle, limited only by exceptions and carveouts. The most important of these are general standards of proportionality and necessity, compliance with data protection law, service security and commercial confidentiality.[96] This broad scope stands to reason since the researchers involved can be held to standards of confidentiality, security and research ethics. Still, the lack of clear limiting principles leaves the question: how sensitive is too sensitive? Even with confidentiality and security guarantees in place, the prospect that researchers might gain unrestricted access, for instance to users' private messages, could be problematic from a privacy perspective. Such important questions are left to subsequent standard-setting in data protection law.[97] And whilst there are no clear carveouts or limiting principles yet for privacy, there are remarkably broad exceptions for service security and the protection of confidential information.[98] These exemptions are phrased broadly and could provide powerful leverage for (notoriously litigious) social media platforms to minimise access and protract access proceedings. In a worst case scenario, the in-depth confidential approach may not offer very much depth at all.

96 DSA, Article 40(2) and 40(5).

97 Standard-setting is already under way to develop a Code of Conduct for access to platform data under the GDPR, which will offer guidelines for researchers on data protection and research ethics. For regulator access there is comparatively less guidance, though this arguably raises the greater privacy concern. European Digital Media Observatory (EDMO), Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access (2022). <https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf> accessed 28 September 2022.

98 DSA, Article 40(2) and 40(5). It is notable, for instance, that the exemption for commercial secrecy protects not only trade secrets as such, but 'confidential information, in particular trade secrets' (DSA, Article 40(2)). Trade secrets are already an exceedingly broad category, and have proven to be a significant barrier to transparency in other fields. This clause is far broader still. Then again, further complicating the matter is Recital 98's proviso that "consideration of the commercial interests of providers should not lead to a refusal to provide access to data necessary for the specific research objective pursuant to a request under this Regulation". This would suggest a more restrictive interpretation of trade secrets and related interests than we have seen in other fields. See, generally, on trade secrets and transparency: Emilia Korkea-aho and Païvi Leino, 'Who owns the information held by EU agencies? Weed killers, commercially sensitive information and transparent and participatory governance' (2017) 57 *Common Market Law Review* 1059.

The CrowdTangle provision, conversely, is rather more limited in the data it offers, and perhaps overly so. Certainly it is vague. Its main target is 'publicly accessible' data, but this category is arguably both under-inclusive and over-inclusive. On the one hand, platforms hold much non-public data which might be amenable to disclosure and valuable to the public. One example from practice is the spending data in platform ad archives; this was not available before the ad archive was implemented, and it has proven to be useful for accountability purposes.[99] In this sense, Article 40(12)'s focus on public content bespeaks a rather limited ambition. Worse, such a rule might even create perverse incentives for platforms, encouraging them to offer *less* data through their public interfaces so that it won't be caught by these new rules.

On the other hand, the category of 'public data' is by itself no guarantee against privacy risks. Article 40(12) emphasises aggregation, which can help to mitigate privacy risks for audience data. But aggregation is less useful when it comes to social media's *content*: when content is aggregated, social media becomes knowable only through the classifiers developed by the platform itself, and observers are unable to develop independent analytical perspectives. The failure of such content-blind approaches can be seen for instance in content moderation reporting, which states aggregate numbers on categories like hate speech or copyright removals but offers no insight into the precise method, impact, or accuracy of these decisions.[100] Comparably, self-regulatory ad archives have been criticised for proffering a selection of 'political' ads detected by the platform, without enabling researchers to study how platforms actually define and enforce this category.[101] When content regulation is at issue, therefore, the most useful disclosures are consistently those which allow access to the underlying content at issue, and do not rely solely on the platform's own classification methods. And yet, this same content can implicate privacy interests even when it is technically public accessible.[102]

The DSA defines 'public' data on a technical basis, as content which is made available to a potentially unlimited number of third parties.[103] But social media's public channels are often used for sensitive and personal interactions, with expectations of privacy stemming not so much from technical accessibility as from practical (algorithmic)

99 Chapter 4 above (Leerssen and others, 'News from the ad archive').

100 Keller and Leerssen, 'Facts and Where to Find Them' (n 74). Bowers, Sedenberg, and Zittrain, 'Platform Accountability through Digital "Poison Cabinets"' (n 75). Common, 'Fear the Reaper' (n 75). Erdos, 'Disclosure, Exposure, and the "Right to be Forgotten" after Google Spain' (n 75).

101 Chapter 3 above (Leerssen and others, 'Platform ad archives: promises and pitfalls').

102 Ben Zimmer, 'Techlash': Whipping Up Criticism of the Top Tech Companies', *The Wall Street Journal* (10 January 2019) <https://www.wsj.com/articles/techlash-whipping-up-criticism-of-the-top-tech-companies-11547146279> accessed 24 September 2022.

103 DSA, Article 3(i).

obscurity.[104] Conversely, technically private groups or channels can take on such a scale and popularity as to attain an effectively public or semi-public function.[105] This is true, for instance, of WhatsApp groups in many countries but also of closed Facebook groups.[106] In this light, social media *entangles* public and private modes of communication, often without a clear dividing line. For Thomas Poell, Sudha Rajagopalan and Anastasia Kavada, social media does not so much create publics localised in specific (virtual) spaces or channels, so much as it gives rise to a dynamic flow of porous and ever-shifting *publicness*.[107] These blurred boundaries pose a serious challenge for observability regulation: solutions which might seem eminently suitable for some items might be highly questionable for others, even within the same technical channels.

In this regard, social media appears somewhat unique. On many other types of platforms, activity tends to be delineated more clearly into two- or multi-sided markets, with private consumers served by public services or retailers. Observability for UberEats menus or Google Play Apps, for instance, is therefore not so fraught with privacy issues as observability for Instagram posts or TikTok videos. On social media the line is harder to draw since ordinary users commonly participate in content creation as well as consumption.[108] And considerations also vary *between* different

104 Danah Boyd, 'Social network sites as networked publics: Affordances, dynamics, and implications', in Zizi Papacharissi (ed.), A Networked Self (Routledge 2010). Woodrow Hartzog and Frederic Stutzman, 'The Case for Online Obscurity' (2013) 101 *California Law Review* 1.

105 It is worth noting that the DSA's definition of platform only covers services which consist in the public dissemination of user-generated content. Whatsapp, therefore, is not a platform for purposes of the DSA. Still, many social media platforms including Facebook and Twitter do offer various channels to conduct private and semi-private communication via their service. See, DSA, Article 3(i).

106 Joelle Swart, Chris Peters and Marcel Broersma, 'Shedding light on the dark social: The connective role of news and journalism in social media communities' (2018) 20 *New Media & Society* 4329. Rafael Evangelista and Fernanda Bruno, 'WhatsApp and political instability in Brazil: targeted messages and political radicalisation' (2019) 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1434> accessed 27 September 2022.

107 Thomas Poell, Sudha Rajagopalan and Anastasia Kavada, 'Publicness on platforms: Tracing the mutual articulation of platform architectures and user practices' . In Zizi Papacharissi (ed.), *A Networked Self and Platforms, Stories, Connections* (Routledge 2018).

108 This blurring of boundaries is already recognised in early Web 2.0 concepts such as the 'prosumer' and 'produsage', although these emphasise the economic or commercial dimension rather than the discursive or civic dimension. One might even say this hybridity is inherent to the social media principle of 'user-generated content', which José van Dijck already described in 2007 as a 'a trade market in potential talents and hopeful pre-professionals', being 'neither exclusively produced amateurs nor by professionals' but rather a 'blending of work and play'. See: Axel Bruns, 'Produsage' (2007) *C&C '07: Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition* 99. José van Dijck, 'Users like you? Theorizing agency in user-generated content' (2009) 31 *Media, Culture & Society* 41.

social media platforms; Twitter, for instance, is a relatively more public medium than, say, Snapchat.[109]

All this means that observability policy for social media needs a thicker concept of publicness than mere technical access can provide. Self-regulation may hint at solutions. For instance, ad archives focus on a specific platform *channel*, namely advertising. This reflects, in essence, the judgement that advertising is and ought to be a public mode of communication. For organic content, some developer APIs employ *threshold values for prominence* or *visibility*, so that certain items are only disclosed for content that has attained a sufficient degree of prominence. And CrowdTangle focuses on specific actors: public pages and 'influential' figures (though it is not clear how they define and enforce these crucial categories; perhaps it is assessed manually on a case-by-case basis).[110] Interestingly, the DSA also refers to 'public figures' as a category subject to Article 40(12), begging the question how this will be operationalised—clearly 'public figures' is already a more complex category than the technical category of 'public content'.[111] Like all line-drawing exercises, developing rubrics for public content will most likely also generate arbitrary edge-cases and exceptions. Still, it will be an essential step towards regulating observability based on a thicker concept of publicness, rather than a simplistic principle of technical access. When—if ever—should (vetted) researchers be able to access (semi-)private groups or messages? Should traffic or engagement data be public for some items, and if so which? When does the public have the right to know when an item has been downranked? Negotiating such questions will require nuanced and creative engagement with the specific affordances of different platforms, and how and when these give rise to publicness worth observing.

In sum, Article 40 DSA's researcher access rules show us the complexity of regulating for observability; it offers few definitive answers, and raises many important questions. The implementation and enforcement stages will be crucial in deciding its role. As we have seen, the DSC of establishment will likely have an appreciable role to play in charting its course; though they lack formal discretion over individual cases, their interpretation, governance and triage of the application procedure could still be decisive to overall outcomes. As to recipients, they might steer Article 40 either toward more in-depth and bespoke projects, decided on a case-by-case basis, or instead

109 European Digital Media Observatory, Report of the Working Group on Platform-to-Researcher Data Access (2022) <https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf> accessed 26 September 2022.

110 Miles, 'What data is CrowdTangle tracking?' (n 63).

111 DSA, Recital 98.

towards a more inclusive, scalable, and automated approach. As to the substance, DSCs and courts together will have substantial leeway in interpreting relevant standards, and must strike a balance between competing interests in observability versus privacy, service security, and platform claims to confidentiality. Overly strict readings could leave even the confidential framework relatively toothless, whilst overly permissive readings of the public framework could jeopardise the privacy of private and semi-public activity such as in direct messages or private group discussions.

Looming in the background of all these issues is the question of purpose: what is it all for, this observability? We have seen that Article 40 DSA represents a relatively specific vision: enforcing the DSA, in particular by contributing to the understanding of 'systemic risks' and their mitigation.[112] I discuss this further below.

4. Regulating social media with observability

Rieder and Hofmann propose observability as a 'companion to regulation'.[113] The DSA also links observability to regulation. And yet their approach differs. Below I discuss in greater detail how observability might in fact contribute to regulation, contrasting the DSA's narrow theory of enforcement against Rieder and Hoffman's broader theory of knowledge production. I then discuss non-regulatory roles for observability in public discourse.

4.1 Observability and regulation

By coupling observability to regulation, Rieder and Hofmann seek to distance observability from transparency's associations with neoliberal deregulation. Rieder and Hofmann, along with many others in the critical transparency literature, argue that disclosure regulation should be conceived of not as an *alternative* to regulation, but as a catalyst or companion. But how, precisely, can observability support regulation?

At an absolute minimum, observability can support regulation by *not detracting* from it. This can occur when transparency is invoked as a cause or pretext against substantive regulation.[114] At the macro-level, the DSA already shows that legislators do not treat transparency and behavioural regulation as strictly either/or proposition; it combines the two. Anecdotally, however, at the micro-level, EU Commissioner Thierry Breton has already invoked the transparency of ad archives as a reason not to enact the Parliament's

112 On systemic risks, see Section 2.1 of this Chapter above.

113 Rieder and Hofmann, 'Toward platform observability' (n 3) 11, 23

114 e.g. Monika Zalnierute, '"Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism' (2021) 8 *Critical Analysis of Law* 39. Pozen, 'Transparency's ideological drift' (n 1).

more far-reaching proposal to ban ad targeting altogether.[115] More in-depth regulatory politics research would be needed to fully grasp the dynamics of such rulemaking practice, but incidents such as these do illustrate the risk of observability being co-opted by deregulatory forces.[116] As a matter of regulatory economy, even well-intentioned observability policies can divert scarce resources away from more robust regulatory interventions.[117] These opportunity costs need not entirely rule out observability as a companion to regulation, of course, but they do counsel for alertness to its rhetorical positioning and its resource allocation vis-à-vis binding behavioural law.

Against these costs, what benefits might observability offer for regulation? Rieder and Hofmann's accounts offers several answers, interrelated but distinct and worth unpacking. The immediate connection is one of *enforcement*, as in the monitoring of compliance: 'whatever set of norms or values are chosen as guiding principles, the question remains how to 'apply' them, that is, how to assess platform behaviour against public interest norms'.[118] In an ideal-typical case, a researcher or journalist might for instance act as a watchdog or 'fire alarm' alerting regulators or other legal institutions to illegality and thereby trigger enforcement action.[119] But observability can also contribute to regulation in more indirect ways, not only in the process of enforcing norms but also in the process of developing norms, through more general mechanisms of knowledge production and public rulemaking. Independent research and reporting about platforms, enabled through observability, can help to guide regulatory institutions such as agencies, legislators and courts in the formulation of these new norms.[120] It might do so directly and through formal channels (for instance through consultations or expert opinions), or indirectly and informally by triggering awareness and debate. This broader knowledge production perspective aligns with Natali Helberger, Jo Pierson and Thomas Poell's model of 'cooperative responsibility', which prioritises exchange between civil society and the state in the defining and operationalising values in platform governance.[121]

115 Kirsten Fiedler, *Twitter.com* (18 December 2020) <https://mobile.twitter.com/Kirst3nF/status/1339889430975410176> accessed 26 September 2022.

116 On the role of regulatory politics research in platform governance, see: Robert Gorwa, 'Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG' (2021) 56 *Telecommunications Policy* 102145.

117 For instance, earlier research into the use of the Facebook Ad Library indicates that a substantial portion of the journalistic coverage on this topic revolved around discussing the Ad Library itself, rather than actually using the data provided by this tool. See: Chapter 4 above (Leerssen and others, 'News from the Ad Archive').

118 Rieder and Hofmann, 'Toward platform observability' (n 3) 23.

119 Peter May, 'Regulatory regimes and accountability' (2007) 1 *Regulation & Governance* 8. Pippa Norris, 'Watchdog journalism', in Mark Bovens, Robert Goodin and Thomas Schillemans (eds.), The Oxford Handbook of Public accountability (Oxford University Press 2014).

120 Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 60).

121 Helberger, Pierson, and Poell, 'From contested to cooperative responsibility' (n 27).

These reflections might seem abstract but they have practical implications for the design and evaluation of observability rules. As to design, we have seen that the DSA's data access framework is predisposed towards enforcement and legal accountability over knowledge production and social accountability. Crucially, the DSA only permits research inquiries related to compliance with its own systemic risk rules. Admittedly these systemic risks are an exceedingly broad category, and can still accommodate a wide range of research questions. Still, the DSA's approach rules out research on novel issues not foreseen by its systemic risk framework, and, indeed, critical research as to the merits of the regulatory apparatus itself—i.e. 'second-order accountability'.[122] Furthermore, it is likely to side-line fundamental and theoretical lines of inquiry, which may only reveal their significance to regulation indirectly and over time, in ways which regulators are ill-placed to predict. In this way, the DSA's enforcement-based approach puts pressure on the principles of academic freedom and free inquiry, per which research agendas are to be determined autonomously by scholarly communities rather than by the external demands of lawmakers or regulators (or, for that matter, platforms).[123] With a greater emphasis on knowledge production and social accountability, observability policy would attach greater importance to such free inquiry, and less importance to immediate enforcement questions. In addition, social accountability perspectives would also attach greater importance to diversity and inclusion, for instance by streamlining, standardising and automating access procedures.

The tension between enforcement and knowledge production theories also recurs in the *evaluation* of observability regimes. Against charges of naïveté and ineffectuality, transparency policies face growing pressure to demonstrate their impact.[124] Legal accountability theories might seem attractive in this context since they focus on relatively concrete and measurable outcomes, such as fines and other regulatory interventions. Social accountability, by contrast, hinges on 'gradual, diffuse, and indirect' effects which are, for Gregory Michener, by their nature 'indirect and challenging to measure'.[125] And yet, legal accountability is not so clearly measurable as

122 Kaminski, 'Understanding Transparency in Algorithmic Accountability' (n 60).

123 Ralph Fuchs, 'Academic freedom – its basic philosophy, function and history' (1963) 28 *Law and Contemporary Problems* 431. Article 13 of the Charter of the European Union also enshrines the freedom of the arts and sciences, including academic freedom—though I do not wish to suggest that Article 40 DSA necessarily violates that right.

124 Baume and Panadopoulos, 'Transparency: from Bentham's inventory of virtuous effects to contemporary evidence-based scepticism' (n 1). Igbal Safarov, Albert Meijer and Stephan Grimmelikhuijsen, 'Utilization of open government data: A systematic literature review of types, conditions, effects and users' (2017) 22 *Information Polity* 1. Maria Cucciniello, Gregory Porombescu and Stephan Grimmelikhuijsen, '25 Years of Transparency Research: Evidence and Future Directions' (2017), 77 *Public Administration Review* 32.

125 Michener 'Gauging the Impact of Transparency Policies' (n 1).

it may seem. A first problem is that independent research can influence enforcement action *informally* and without leaving a trace on the official enforcement record. Indeed, from an ecological perspective on transparency, any individual item of research can have many unforeseen knock-on effects on other research activities, challenging our capacity to isolate the impact of any specific act.[126] A second problem is deterrence: watchdog activity can discipline platforms into compliance even in the absence of actual enforcement actions—so long as it raises the *perceived risk* of such enforcement action. Such instances of platforms adjusting their policy in response to public criticism are well-documented.[127] By extension, we can postulate that observability *itself* can discipline platform conduct by its very presence, even in the absence of actual watchdog usage—so long as it raises the *perceived risk* of such watchdog activity occurring. Such deterrence effects, or the 'potentiality' of legal accountability, are counterfactual in nature and largely beyond direct measurement.[128] All this means that observability's effects on legal accountability are scarcely any more straightforward to measure than its contributions to social accountability; indeed, in practice these legal and social mechanisms are inextricably linked. A knowledge production perspective, I submit, leans into these measurement issues by treating independent research and monitoring as a goal in itself; knowledge production not merely a means to the end of legal sanctions (though it may well support this end!), but as a presumptive good by its own merits.[129]

In sum, the above reflections all speak to the basic concern that observability, through an overly tight coupling with regulation, risks being flattened from its initial conception as a broad instrument of social accountability to a thin instrument of legal accountability. Such an emphasis on legal accountability and enforcement might appear hard-nosed and realistic, an antidote to the naïveté of transparencies past, but it risks ignoring much of what makes observability important to regulation and to democracy. Though it would certainly be naïve to treat data access as a self-

126 Kreimer, 'The freedom of information act and the ecology of transparency' (n 78).

127 Chapter 4 above (Leerssen and others, 'News from the ad archive'). Bossetta, 'Scandalous design' (n 61). Bridget Barrett and Daniel Kreiss, 'Platform transience: changes in Facebook's policies, procedures, and affordances in global electoral politics' 8(4) *Internet Policy Review* < https://doi.org/10.14763/2019.4.1446> accessed 25 September 2022.

128 Richard Mulgan, 'Accountability: An Ever-Expanding Concept?' (2000) 78 *Public Administration* 555. Albert Meijer, 'Transparency', in Mark Bovens, Robert Goodin and Thomas Schillemans (eds.), *The Oxford Handbook of Public Accountability* (Oxford University Press 2014).

129 Chapter 4 above (Leerssen and others, 'News from the ad archive'). Katharine Dommett, 'The inter-institutional impact of digital platform companies on democracy: A case study of the UK media's digital campaigning coverage' (2021) *New Media & Society* <https://doi.org/10.1177/14614448211028546> accessed 25 September 2022. Rui Pedro Lourenço, 'Evidence of an open government data portal impact on the public sphere' (2016) 12 *International Journal of Electronic Government Research* 21.

executing policy panacea, it is no less naïve to reduce its complex social and political ramifications solely to their legal outcomes. The best defence of observability requires a more fulsome appreciation of knowledge production as a means to democratic self-governance. From this perspective, observability policy might still be oriented towards regulation, but rather as a general north-star principle than as an immediate objective or deliverable.

4.2 Observability and discourse

Observability affects public discourse. Previously I mentioned public deliberation as part of the regulatory 'social accountability' of platforms, but what I'm referring to here is broader than that, and specific to social media platforms. Namely: social media platforms are not just *topics* or *objects* of public deliberation and accountability, but also *sites* and *mediators* of public deliberation. The same civil society actors who might avail themselves of observability tools often participate in these online media discourses which they study, and might even be active users of the very platforms they specialise in. Observability, then, has the potential to work back on the publics which it documents, and alter the terms of engagement within and between them, in ways which are not necessarily, solely or even primarily matters of legal or regulatory concern. Observability, in other words, can be understood not only as a barrier to regulation or accountability but as an affordance of social media technologies which shapes the communicative process itself.

An example: Journalists have used ad archives for the conventional watchdog work of uncovering wrongdoing, but also simply to report on campaign messaging strategies which were previously off the record.[130] Political campaigners too have used ad archives in order to study and respond more effectively to their opponents' messages and campaign strategies.[131] If personalised campaigning contributes to the 'fragmentation' of the public discourse, as Tom Dobber, Ronan Fahy and Frederik Zuiderveen Borgesius argue, then this type of observability seems to go some way in *defragmenting* it.[132] Similarly, CrowdTangle has been used by journalists to highlight trends in online right-wing extremism, and this can be understood as valuable not only as a prelude to *regulating* these speakers but also as a tool for media criticism; for reflection and

130 Chapter 4 above (Leerssen and others, 'News from the ad archive'). Bossetta, 'Scandalous design' (n 61).

131 Bossetta, 'Scandalous design' (n 61).

132 Tom Dobber, Ronan Ó Fathaigh and Frederik Zuiderveen Borgesius, 'The regulation of online political micro-targeting in Europe', 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1440> accessed 24 September 2022. Frederik Zuiderveen Borgesius and others, 'Online Political Microtargeting: Promises and Threats for Democracy' (2018) 14 *Utrecht Law Review* 82.

discussion about and across these divides hewn by algorithmic personalisation.[133] In this way, observability aligns with the ideal of the public sphere as a reflexive social space, where conduct is coordinated in the first instance through discursive exchange rather than through legal obligation; a 'space of discourse organised by discourse'.[134] To this end, publics must be able to witness themselves; to engage in what Isabel Kusche terms the 'self-observation' of public discourse.[135]

This point is related to, but distinct, from the common critique of personalised social media as creating 'filter bubbles' or 'echo chambers'.[136] Such theories describe a first-order problem of media diversity in personalised flows, whereas I am describing a second-order problem in the observability of those personalised flows. If the filter bubble theory speaks to the algorithmic subdivision or fragmentation of the public sphere into insular 'sphericules', then observability critiques the capacity for participants to *identify* these sphericules and *locate* them in relation to each other.[137] In some cases, the second-order programme of observability regulation may be a precursor to the first-order programme of recommender regulation. But in other cases, observability regulation may act by itself as an affordance for more engaged and cohesive online discourses. If the filter bubble theory invites us to 'pop the bubble' by reversing personalisation and mandating standardised offerings from social media, then observability invites data access which helps communication to flow reflexively and deliberately amongst these algorithmic publics. Observability does not demand that we all see the same content, but simply the capacity to see what others are seeing.

133 Richard Rogers, 'Social media research after the fake news debacle' (2018) 11 *Partecipazione e conflitto* 557. Kevin Roose, 'Inside Facebook's Data Wars', *The New York Times* (14 July 2021). <https://www.nytimes.com/2021/07/14/technology/facebook-data.html> accessed 6 November 2022.

134 Michael Warner, *Publics and counterpublics* (Princeton University Press 2021) 68.

135 Kusche also refers to this as 'second order observation', a fitting companion to Kaminski's 'second order accountability' in the specific context of media governance and its governance of public speech. Whereas second-order accountability speaks to the observation of governance arrangements, second-order observation speaks to the observation of discourses. Isabel Kusche, 'Private Voting, Public Opinion and Political Uncertainty in the Age of Social Media' (2022) 51 *Zeitschrift für Soziologie* 83. See also: Joelle Swart, Chris Peters and Marcel Broersma, 'Shedding light on the dark social: The connective role of news and journalism in social media communities' (2018) 20 *New Media & Society* 4329. Warner, *Publics and counterpublics*.

136 Eli Pariser, *The Filter Bubble: What the internet is hiding from you* (Penguin 2011). Axel Bruns, *Are Filter Bubbles Real?* (Polity Press 2019).

137 Todd Gitlin, 'Public sphere or public sphericules?' in James Curran and Tamar Liebes (eds), *Media Ritual and Identity* (Routledge 1998). For more recent commentary in this vein, see: Philip Schlesinger, 'After the post-public sphere', (2020) 42 *Media, Culture & Society* 1545. Stephen Coleman and Karen Ross, *The Media and the public* (Wiley 2010), Ch. 6.

5. Conclusion

Observability invites us to think through the challenges of platform transparency in new ways. This paper has taken up that invitation, and shown how this pragmatic, sociotechnical perspective can inform ongoing legislation in Europe. With the DSA, platform regulation is taking a major leap, and observability its first small steps. In doing so it presses us to refine our theory of observability and its relationship to the broader regulatory push. In this paper I have approached this relationship along its two axes; regulating for observability, and regulating with observability. By focusing my discussion on social media and their recommender systems, I have also tried to show how observability might matter differently in one specific and especially controversial domain of platform governance.

The DSA's model for observability regulation is ambitious, and so open-ended in its standards that predicting its requirements remains difficult. Its initial design already surfaces important challenges. Its central dilemma is that of depth versus scale and inclusivity: should it provide high-level data to many, or in-depth data to few? Its current design seems to attempt both, at least on paper, regulating both confidential data sharing agreements and automated, scalable interfaces. Its trajectory in practice will depend largely on how the DSC of establishment governs application procedures, and, together with courts, interprets relevant exceptions. One particular challenge for more scalable approaches, especially for social media, is developing rules that start to disentangle public from private communication, and thus to start establishing reasonable expectations of observability.

As for observability's contributions to regulation, I have shown how the DSA risks flattening this principle from a broad instrument of social accountability and knowledge production into a more narrow instrument of legal accountability and enforcement. Realising observability's full potential, I have argued, calls for a looser coupling with regulation, attentive to the many informal and indirect ways that knowledge production can support regulation; it calls for a greater emphasis on inclusiveness and free inquiry than the DSA currently suggests. And for social media, I have argued, observability's role extends beyond regulation and into discourse, providing new opportunities for exchange and deliberation across and about social media's fractured publics.

CHAPTER 7

Conclusions

All of the verbs for public agency are verbs for private reading, transposed upward to the aggregate of readers.

Michael Warner, *Publics and Counterpublics* (2002)

1. Introduction

In early 2016, a photograph of Facebook CEO Mark Zuckerberg went viral, spreading across magazines and newsfeeds around the globe and prompting countless memes and parodies.[1] Taken at the Mobile World Congress, the image depicts a large audience trying out a demo for Facebook's new virtual reality (VR) headsets. We see hefty black boxes strapped to everyone's face, covering their eyes and absorbing them in a personal virtual experience. Overseeing all this, the only person without a mask, is Zuckerberg himself.

Many took the image as a portent of VR's dystopian qualities, which promised a 'horrible cyberpunk future'.[2] But in some sense this future is already present: with or without VR, it plays out every day on social media. Each individual experiences a personalised flow of curated content, without the capacity to see what others are seeing. Only the platform sees across these flows, and their power to shape this process is hidden from public view.

In this dissertation, across five papers, I have studied legal responses to this problem. How can EU law regulate the transparency of recommender systems in order to hold online platforms accountable for their role in social media governance? My inquiry was guided by the following sub-questions:

1. What different models of accountability are reflected in the EU's regulation of recommender system transparency for social media?
2. How can transparency regulation contribute to the accountability of content curation through recommender systems?
3. How can transparency regulation contribute to the accountability of content moderation through recommender systems?
4. What is the relationship between transparency regulation and behavioural regulation in social media recommender governance?

This final chapter reviews my findings. It proceeds in two parts, corresponding to the core characteristics of transparency regimes: substance (transparency of what?) and audience (transparency for whom?). Along these parallel tracks, I discuss my answers to the above sub-questions. On this basis, I then answer my main research question. I close with an outlook on future challenges for legal and interdisciplinary scholarship.

1 Rick McCormick, 'This image of Mark Zuckerberg says so much about our future', *The Verge* (22 February 2022) <https://www.theverge.com/2016/2/22/11087890/mark-zuckerberg-mwc-picture-future-samsung> accessed 28 September 2022.

2 Ibid.

2. Transparency of what? From algorithms to sociotechnical systems

The law has come to recognise platform recommender systems as important points of control in social media governance. This concern is reflected in various proposals across Europe to regulate them, in particular by making them more transparent. Chapter 2 commenced my analysis by reviewing these proposals in their current state of play. Here we saw that transparency rulemaking often approaches recommender transparency in terms of *algorithms*, where the main goal is to 'open the black box' and *explain* the algorithmic logics behind their automated decisions. In this project I have tried to broaden that perspective.

With the help of social media studies and algorithm studies, we have seen that recommender systems are not reducible to a single monolithic algorithm, and are better understood as dynamic sociotechnical systems comprised of a concatenation of human and computational actors. This perspective is founded on the basic premise of STS that technology's societal effects are not inherent to artefacts but co-determined by their perception and usage; their embeddedness in social practice. This is true for all technology, and it is especially salient for social media recommender systems, being not only technically complex but also highly interactive and user-driven in their performance. Current approaches which focus on *algorithmic* transparency therefore risk overlooking this crucial context of usage, this social embeddedness. Audience engagement patterns, uploader content markets, advertiser targeting strategies, platform interventions—all key factors in the recommending process—change over time, across populations, and in response to the algorithm and each other. A sociotechnical perspective, therefore, brings us from a singular concern with recommender algorithms to a more general engagement with recommender systems and their usage; not only as technical artefacts but as sites of coordination and governance, through which platforms regulate behaviour on their services.

A sociotechnical perspective, therefore, is at the heart of my subsequent analysis of recommender systems, in terms of the different governance functions they fulfil for platforms: content curation and moderation.[3] As means of curation, recommender systems regulate content by defining relevance and producing visibility, which is achieved through an iterative and interactive process of engagement optimisation.

3 This distinction is drawn the work of Poell, Nieborg and Duffey in *Platforms and Cultural Production*, which appeared only after the publication of Chapter 2. A similar distinction is present in Chapter 2, albeit in slightly different terminology: here I contrast fundamental or systemic changes (i.e. curation) against content-specific or targeted downranking decisions (i.e. moderation).

As a means of moderation, recommender systems regulate content by restricting visibility for specific items, which is achieved through automated or semi-automated industrial processes to screen and classify user content and behaviour. As distinct modes of governance, curation and moderation also call for different approaches to transparency and accountability, to which I dedicate separate sub-questions and chapters.

Understood in terms of content curation and moderation, recommender system transparency reveals other salient questions besides conventional algorithmic explanations. For curation, we must consider inputs (How are users engaging with, uploading and targeting content?) and outputs (What is being recommended and to whom?). For moderation, we must be able to observe how the platform intervenes in recommender systems to sanction and demote content (What is being demoted or delisted?). Observing these aspects is presently all but impossible not only due to platform secrecy but also due to personalisation; since every user receives personalised recommendations, understanding the recommender at a systemic level is challenging. In this sense, the issue of algorithmic explainability, and the black box metaphor with which it is associated, do not go far enough in describing the opacity of recommender systems. The following chapters offered more detailed case studies on disclosure models for each mode, and how they respond to this problem. Chapters 3 and 4 focused on ad archives as windows onto algorithmic curation (addressing sub-question 2), and Chapter 5 on due process rights as a window onto content moderation interventions (addressing sub-question 3).

Ad archives, introduced in Chapter 3, are a novel disclosure model in social media governance which illustrate a sociotechnical approach to algorithmic content curation. These tools create a public record of personalised communication by documenting the content, buyer identity, and audience demographics of platform advertising flows. In doing so, ad archives open up personalised curation to public accountability, enabling outside scrutiny of these algorithmically-segmented discourses and of the platforms' responsibility in curating them. During the time of writing my analysis, these ad archives were largely unregulated by law (with the exception of Canada and the State of Washington). Since then the DSA has made ad archives a binding requirement for all large platforms in the EU.[4] Reviewing criticism from advertising researchers, this study has highlighted three important criticisms of extant self-regulatory offerings—scoping, verifying and targeting—which underscore the importance of binding regulation for these tools.

4 DSA, Article 39.

These pitfalls in ad archive governance speak to some important challenges for transparency in content curation. First, the matter of scoping—what qualifies as a 'political' advertisement?—highlights the limited capacity of platforms to reliably categorise content. As a consequence, disclosure rules based on such categories of interest may be inadequate, because they merely reproduce those very platform classification methods which are commonly at the heart of regulatory controversy, and provide no basis to critique these methods or examine what the platforms might be overlooking. More holistic access based on functional, technical categories—such as archives covering *all* ads—are preferable since they allow third parties to develop their own analytical perspectives, and to independently critique those applied by platforms. Second, the matter of verifying speaks to the risk of overreliance on platform data as an objective source of truth; here too, effective transparency regulation should be attentive to the limited control and knowledge that platforms exercise over their services, and the potential for deception or obfuscation by their users. In some cases, effective transparency may therefore require verification duties for platforms to ensure the reliability of the (meta)data they disclose, such as ad buyer identities.[5] Third, the matter of targeting highlights user privacy as an obstacle for algorithmic explanations. Audience analytics can be aggregated to prevent privacy infringements, but the appropriate level of detail remains a point of contention. And since the targeting mechanisms used by advertisers may also rely on personal data, fully documenting these mechanisms (i.e. algorithmic logics) is especially fraught. In this sense, ad archives illustrate how an emphasis on outputs (i.e. who has seen the ad?) and inputs (i.e. what targeting instructions did the ad buyer select?), can provide meaningful insights into algorithmic content curation without necessarily insisting on more cumbersome and fraught algorithmic explanations.

Chapter 4 conducted an empirical test of ad archive transparency. It focused mainly on the audiences and accountability functions, which I return to in the second half of this conclusion. As to the substance of disclosures, this study does underscore the continued importance of data scraping, since the most specialised journalists I spoke to still relied on scraping to enrich and indeed verify ad archive data. Furthermore, the content analysis indicates that political advertising is a far more salient issue in the US than in the Netherlands and Germany, and presumably most other EU countries where

5 The DSA's final amendments added a verification duty to its ad archive provision, requiring 'reasonable efforts' to ensure accurate and complete information (DSA, Article 39(1)). Verification has also become a point of discussion in the recently-proposed political advertising regulation. See: Max van Drunen and others, 'Transparency and (no) more in the Political Advertising Regulation', *Internet Policy Review* (25 January 2022) <https://policyreview.info/articles/news/transparency-and-no-more-political-advertising-regulation/1616> accessed 6 November 2022.

platform advertising is less prevalent. Still, ad archives may remain an important window onto commercial advertising in these countries. More fundamentally, this finding reminds us of the centrality of *organic* content on social media, and challenges us to explore how similar disclosure models can be extended to study this domain—a question I return to further below.

Chapter 5 examined the transparency of content moderation via recommender systems; how platforms intervene in the ranking process through demotion or visibility reductions, in order to enforce rules on content and conduct (sub-question 3). I have argued that these new visibility reduction strategies in content moderation are qualitatively different from conventional takedown sanctions which have historically preoccupied legal debates, in that visibility sanctions are uniquely opaque. Whereas conventional takedowns are mostly opaque in their *reasons*, demotion is also opaque in its *outputs*, since it plays out through volatile and personalised recommender systems which work to obscure their impacts. This is reflected in current discourse on 'shadow banning', which expresses pervasive anxieties about the invisible threat of demotion.

Legal debates about moderation transparency have until now largely focused on providing *explanations* or *reasons* for algorithmic decisions, but this Chapter has highlighted the primacy of notification as a prior, minimal safeguard against shadow banning. The DSA, through Article 17's Statement of Reasons, attempts to regulate both notice and explanation for moderation decisions, and in this way offers a bulwark against shadow banning. Should platforms raise security- or cost-based objections to this provision, I have suggested that, in future, a more nuanced balance might be achieved by unbundling light-touch notice safeguards from more burdensome explanation duties. An outstanding question, which I return to below, is whether information about visibility reductions should also be made known to other actors beside the affected uploader, who remains central in the DSA's due process model. Finally, in this chapter we also saw how the goals of curation and moderation transparency are interrelated: shadow banning safeguards for moderation may be difficult to enforce precisely because, at present, overall curation outcomes are still opaque, to uploaders as well as to the public. Making curation transparent may therefore also help to observe the impacts of moderation in these systems, and enforce due process more effectively.

Finally, Chapter 6 scrutinised the language of transparency itself and how the concept of observability, proposed by Bernhard Rieder and Jeannette Hofmann, can help to articulate a sociotechnical approach to recommender transparency. In addition to Rieder and Hoffman's pragmatism, I highlight observability's *decentered directionality*

as an implicit rejoinder to the metaphor of 'opening the black box'. The DSA, I show, contains several provisions which start to regulate observability: its provisions on ad archives and moderation archives but most significantly its general framework for researcher access. This framework, we have seen, raises difficult questions as to the substance of disclosure. One of its key organising principles is whether the data in question are publicly accessible. But even public content on social media platforms can implicate privacy interests of its own, owing to the porous and ever-shifting boundaries between public and private communications on these services. Regulating observability for social media, more so than other types of platforms, therefore requires rubrics to start disentangling public from private speech, based on factors such as the channels used, the actors involved and the visibility attained. Beyond this, my main criticism of the DSA's access framework is one of purpose and regulatory politics, which I discuss below.

3. Transparency for whom? Regulating disclosure for cooperative responsibility

Transparency should consider its audience. For its format, presentation, and process, transparency policy must decide what stakeholders it means to address. In doing so, transparency reflects a regulatory politics; an accountability ideal which includes some and excludes others. In this dissertation I have inquired into these regulatory politics for recommender transparency, and asked how EU law can best reflect a model of cooperative responsibility: an inclusive approach that emphasises interaction between individual users, government regulators and civil society.

Chapter 2, having reviewed various transparency proposals found in recent EU policymaking, inquired into the types of accountability relationships these rules pursue (sub-question 1): user choice, public ordering and independent research. From a media policy perspective, both user choice and public ordering are problematic—if not irrelevant, then at least insufficient by themselves. User choice does not guarantee the realisation of public interest principles such as quality or pluralism, and direct public ordering threatens media freedom through concentrations of systemic opinion power. Precisely if governments are to regulate platform recommenders (by law or otherwise), then transparency of these systems becomes all the more salient in order to ensure second-order accountability of the resulting hybrid power arrangements. This holds true especially for social media governance, where technocratic appeals to neutral or objective expertise are unlikely to resolve the inherently political conflicts at stake. These considerations highlight the importance of cooperative responsibility

in social media governance, and, accordingly, *inclusive* transparency which involves not only users or governments but also civil society and the public.

Recent policymaking, I have shown, has started to explicitly address such demands for civil society transparency. And yet, the dominant approach is one of selective, confidential access for vetted researchers, which aims to manage the privacy and security risks associated with disclosing platform data by limiting the personal scope of access. But these advantages must be weighed against important downsides, I have argued: vetting and accreditation procedures come at a cost to the scalability and inclusiveness. The data's potential reach and uptake are restricted, and especially non-academic civil society stakeholders such as journalists, activists and political actors are likely to be marginalised by such an approach. The EU's model for civil society access therefore tends towards an exclusive, expert-driven, technocratic approach rather than a more inclusive and overtly political one. Going against this trend, this project has therefore tried to articulate the distinct advantages of public transparency resources.

Chapters 3 and 4 considered the role of such a public resource, platform ad archives, and how these can contribute to accountability of content curation (sub-question 2). Precisely because their public design does not prefigure a specific purpose or use case, I have tried to articulate one. Across two papers, I have analysed and tested the various forms of legal and social accountability which ad archives can enable, and the interactions between them. Conventional accounts of transparency and regulation may focus on 'fire alarm' scenarios, where civil society actors detect wrongdoing and trigger legal repercussions. But we have seen that this account can be expanded in several ways in the context of social media and its content curation. First, in platform governance, our understanding of 'legal' accountability must recognise that many important norms are regulated by platform design choices and policies rather than public law. Second, platforms can also be responsive, at least under certain conditions, to public criticism and deliberation, including wrongdoing which is not formally circumscribed in any way. Third, ad archives and other data access resources can also be understood as influencing the structure of public discourse itself; by providing participants the opportunity to observe and respond to personalised communication flows. For all these reasons, the accountability function of content curation transparency should not be reduced to one of mere legal enforcement, and should also acknowledge its indirect, social and second-order accountability functions.

As for content moderation (sub-question 3), Chapter 5 found that the DSA's transparency rights are designed for an entirely different form of accountability: individual due process. Their basic function is different: not instrumental but justificatory. Due process

rights such as the DSA's do not necessarily aim to factually describe platform conduct in any systemic fashion so much as they aid in the establishment and vindication of users' individual rights vis-à-vis these services. Due process transparency forces platforms to codify their rules and justify individual decisions on this basis, through procedures for notices and appeals. It negotiates tensions with service security by selecting out specific 'bad actors' (commercial spammers) who are ineligible for disclosure. This procedural legal accountability for content moderation actions is one of the DSA's principal objectives. And yet, although due process is primarily designed for the protection of individual user rights, we have seen that in the specific context of recommender systems it can have important interactions with broader forms of collective resistance and social accountability. This is due to the output-level opacity of recommender systems and their visibility reduction techniques, which results in shadow banning, and thereby hides moderation actions not only from the affected user but also from their audiences and from the public at large. Conventional moderation occurs in the public eye, but shadow banning does not. Consequently, the individual safeguards produced by the DSA can still be of some collective significance if and when the affected user decides to make the shadow banning decisions known to the public; in this way, individual due process transparency feeds into public transparency, and individual due process accountability feeds into collective and social accountability. Going further, an important question for future policymaking is under what conditions other stakeholders and the public at large might have an independent right to know about content moderation actions, for instance through audience-facing disclaimers, confidential archives or public databases.

Chapter 6 drew together these findings on the accountability functions of recommender transparency, and reflected on their relationship to behavioural regulation in cooperative responsibility (sub-question 4). It did so by critiquing the DSA's researcher access framework, showing how its design lays bare the tensions between observability as a means to knowledge production and as a means to regulatory enforcement. The proposal is relatively rigid in its insistence on serving only pre-defined categories of researchers and only for the purposes of monitoring compliance with the DSA's own rules. In this way, the DSA couples transparency tightly to regulatory enforcement. It risks marginalising more open-ended and fundamental research as well as more overtly political and deliberative goals. Accommodating these will require a change in our understanding of transparency's regulatory function; not just as an immediate instrument of enforcement but as a more general resource for knowledge production and public discourse. Observability ought to act as a companion to regulation, but the DSA treats it more like a deputy sheriff.

4. Conclusion

In this project, I have asked how EU law can regulate the transparency of recommender systems in order to hold online platforms accountable for their role in social media governance. I conclude that transparency policy under EU law should be *inclusive* in its audience and *sociotechnical* in its substance—twin principles which can be articulated jointly through the concept of *observability*.

An inclusive approach to transparency follows from my normative commitment to cooperative responsibility as a model for platform governance. This model does not seek to affix responsibility in any single actor, such as the platform, the user, or government, but instead aims to facilitate dynamic interaction between these actors. Accordingly, cooperative responsibility does not call for just one single form of transparency, but instead for a variety of disclosures aimed at the specific needs and interests of different actors. For users, these may be simplified, digestible explanations or notices. For regulators and researchers, more in-depth and sensitive data access is warranted. Such a variegated approach to transparency is, as such, not necessarily new in the academic literature on platform transparency. However, recent policymaking discussed in this dissertation, such as the DSA, does represent one of the first major attempts to realise it in practice, especially with its novel focus on civil society access. In this context, the main aim of my dissertation has been to unpack how transparency policy institutionalises 'civil society' as a third category between platforms and regulators, privileging some actors whilst excluding others, and often foregoing fully public resources out of concern for privacy or security.

At its most general level, therefore, this dissertation's recommendation is for law and policymakers to consider seriously the value of inclusive and broadly accessible transparency resources in social media governance. They risk being undervalued, since they are not directed toward any obvious, pre-identified audience or purpose, and pose relatively high risks to privacy and security; both of these factors, it seems, lead policymakers to gravitate towards more targeted and exclusive arrangements instead. And yet public and inclusive resources, I have argued, are not only more scalable in their impacts but also fulfil a distinct role in creating *social accountability* of platforms and their users as well as *second-order accountability* of the governance system itself; opening it up to more overtly political and non-institutional actors such as political figures, activists and journalists who play an important role in public discourse and deliberation. The added value of this social and second-order accountability is especially significant in the context of media ecosystems, where a basic publicity of outcomes has historically been a structural affordance of the

ecosystem and where direct government intervention is especially fraught from a freedom of expression perspective.

What types of inclusive transparency are possible in the context of recommender systems? I have argued for a sociotechnical perspective, which views platform recommendations not only in terms of algorithms but as a product of complex and dynamic interactions between algorithms, their users, and their operators. From this perspective, we can start to inquire into the governance function of recommender systems not only by virtue of their technical characteristics but by asking how this technology is used in practice. Accordingly, the middle section of this dissertation explored case studies for a sociotechnical approach, organised around two different modes of governance exercised fulfilled by recommender systems: content curation, and content moderation. For content curation, a sociotechnical perspective expands our attention from the algorithmic logics of engagement optimisation as such toward the specific outcomes being realised in practice through the algorithm's interactions with users. Prior to the issue of *why*, we must ask: *what* attracts engagement and *what* is being recommended? Likewise, for content moderation, a sociotechnical perspective demands information as to practical impact of demotion in specific cases. Again, prior to the issue of *why*, we must ask: *what* is being demoted? Across both case studies, we have seen that personalisation obscures these basic outcomes, and transparency reforms are necessary to render them observable. I do not argue that these reforms are a sufficient alternative to explanation, but rather that they are an essential first step toward explanation—and, crucially, they are forms of transparency which can be made known broadly, compared to the sensitive and complex data at stake in explanation.

The principle of observability, which I draw from the work of Bernhard Rieder and Jeanette Hofmann, expresses both the multistakeholder and sociotechnical aspects of this approach. This pragmatic concept highlights the differing perspectives and capacities of different observers, and thereby comports with the variety of different stakeholders in platform governance. In particular, by emphasising knowledge production as a key target, Rieder and Hoffman highlight the role of civil society actors in achieving cooperative responsibility. And through its decentered perspective, I argue, observability looks beyond the black box metaphor and the associated paradigm of algorithmic explainability in order to accommodate a more fully sociotechnical perspective.

With this new programme of observability regulation comes the need to revisit transparency's relationship to other regulatory projects. How and when can we consider observability regulation to have achieved its goals, and, from a legal perspective, to what extent must it be accompanied by behavioural regulation? On this

issue, this dissertation has tried to strike a balance between two competing narratives in recent scholarship. On the one hand, the current literature on transparency pushes for a sceptical and evidence-based approach which treats transparency primarily as an instrument for the enforcement of binding regulations; a backlash, one might say, against earlier myth-making about attentive publics, armchair auditors and self-executing transparency panaceas. These lessons are highly relevant to platform governance, since dominant platforms are often capable of acting with little regard for the 'soft' accountabilities of commercial or social pressure, absent any 'hard' threat of regulation to back it up. And yet, on the other hand, social media governance is typified by unprecedented information asymmetries which frustrate independent knowledge production, even more so than in conventional mass media, and it would be simplistic to focus only on the legal ramifications of this societal shift. Furthermore, social media implicates important freedom of expression concerns which militate against integral ordering by public law, and for the involvement for civil society.

Under these circumstances, the relationship between transparency and behavioural regulation deserves a nuanced treatment. The credible threat of legal or regulatory repercussions may often be essential as a means to hold platforms to account, and one of transparency's main goals should be to facilitate this. This renders suspect any rhetoric which invokes transparency as a sufficient alternative to regulation. It should also instil vigilance against *excessive* investments in transparency policy, if and when it imposes significant opportunity costs on more far-reaching behavioural policy. At the same time, a nuanced appreciation of transparency's regulatory functions should at least acknowledge the complex and indirect interactions between social and legal forms of accountability, which have been a recurring theme in this dissertation. Ad archives can facilitate litigation but also voluntary takedowns, public criticism, or more responsive political communication and fact-checking. Moderation notices can trigger individual appeals but also collective outrage or platform switching. For all these reasons, I have argued for a loose coupling between transparency and behavioural regulation, which treats data access not only as a means of enforcement but also as a means of knowledge production; a regulatory politics which embraces transparency not only for the sake of forensics but also for the sake of informed and inclusive deliberation.

5. Outlook

What does the future hold for the observability of social media recommender systems? I will close with an overview of legal and interdisciplinary challenges.

5.1 Legal challenges

The DSA takes important first steps for observability regulation, but also leaves hurdles ahead and uncharted territory beyond. Its rules on ad archives and content moderation due process are, as discussed, important windows onto recommender system governance. Still, the DSA leaves other questions open. First, it still contains many open standards leaving room for interpretation. (To name only a few: what are the 'main parameters' used for ad targeting; the scope of spam exceptions for moderation Statement of Reasons; or the meaning of proportionality for data access applications.[6]) In implementing these open standards, legal institutions will need to weigh transparency against competing interests in user privacy, service security, and platforms' claims to proprietary business interests.

Second, the DSA leaves other aspects of recommender observability untouched, in both content moderation and curation. For curation, the DSA makes much progress on advertising content but leaves largely unaddressed the far larger issue of organic content. Regulating observability in organic content is admittedly far more problematic, since much of it implicates non-public or semi-public communication with privacy interests at stake. Still, existing self-regulatory practices, discussed in Chapter 6, already point to some possible avenues, including tools such as CrowdTangle, which focus on specific subsets of public actors, channels and items. The DSA, however, does not offer specific solutions here. For organic curation it only specifies a general algorithmic explanation rule, which falls short of true observability.[7]

For moderation, an important question remains whether and how individual visibility reductions ought to be made known to broader audiences. In keeping with the DSA's due process model, it is primarily concerned with providing individual redress to affected uploader. Therefore, audiences, regulators and researchers will only be able to observe these matters insofar as the affected user makes them known. This issue is salient to all moderation actions but especially so for visibility restrictions, since they leave almost no observable trace. In some cases there may be privacy rights or security interests at stake in keeping moderation actions secret (or, more specifically, in keeping them between the platform and the uploader), but these interests in secrecy must be weighed against the public interest in observing platform gatekeeping. The DSA's rules on moderation reporting and archiving unfortunately fall short of providing decision-level insights, and it remains an open question for future policymaking whether and how this might still be required.

6 DSA, Articles 39(2)(b), 17(2) and 40(8)(e).

7 DSA, Article 27.

For both content curation and moderation, the general data access framework of Article 40 DSA may provide starting points for regulators to start addressing these open questions. In the hands of well-equipped and well-motivated regulators, the DSA's data access frameworks could perhaps be used to push for inclusive and scalable observability solutions in line with the above, such as Crowdtangle-like traffic dashboards or APIs, or content moderation archives. Here I see few grounds for optimism, however. The statute itself is held back by far-reaching exceptions, as well as the restrictive scope of research purposes and eligible recipients. And its implementation relies heavily on the initiative of capacity-constrained national regulators, and in particular the Irish Digital Services Coordinator, which will have jurisdiction over many of the largest social media platforms. If these authorities will take any steps at all to regulate observability, they may well opt for more risk-avoidant and piecemeal access grants rather than more ambitious scalable and inclusive automated solutions. From a strategic perspective, it may indeed be a sensible choice to start at first with such smaller-scale confidential experiments before scaling up and out to more inclusive solutions. But it is my hope that transparency policy does not lose sight of these greater ambitions. The DSA takes important first steps, but observability regulation still has a long road ahead.

5.2 Interdisciplinary challenges

Looking beyond law, this dissertation's findings point to the growing importance of integrating legal scholarship on social media transparency with empirical work in social media studies. Their reliance is mutual. In one direction, we have seen that empirical work in the platform society faces complex legal challenges, and, now with the DSA, new opportunities. Researchers wishing to study platforms must therefore learn to navigate a thicket of legal strictures, from Terms of Service contracts to GDPR Codes of Conduct and DSA access requests. In the process, they will be forced to negotiate the regulatory politics of these legal frameworks: which rules to comply with and which to transgress; how to negotiate, through transparency law, the shaping of research agendas by platforms and regulators; how to mitigate threats to academic and intellectual freedom. To this end, more than ever, social media research needs social media lawyers.

In the other direction, lawyers and policymakers must engage with empirical researchers in order to understand transparency's effects in practice. Transparency has failed too often to be taken on faith, and its design should therefore be based on firm evidence of usage and responsiveness to its audience(s). Effective transparency policy therefore demands an empirical research agenda of its own: only through this work can we hope to develop transparency policy which meets the needs of its

audiences, and strikes an appropriate balance with competing interests. At the same time, as I have argued throughout this dissertation, the push to develop a clear-eyed, non-naïve, and evidence-based view of transparency should not lead us to flatten or oversimplify its potential. Transparency's effects do not manifest solely in spectacular 'smoking guns' or 'fire alarms', but also in more diffuse processes of public awareness, deliberation, and knowledge production, which can be more difficult to observe directly. A more nuanced appreciation should also accommodate these subtler forms of social and second-order accountability, which may be non-obvious to lawyers especially. This entails the careful reflexive work of studying the reception and usage of transparency and observability tools by different stakeholders— research about platform journalism, research about platform activism, and, indeed, research about platform research. Even if transparency is no longer a 'quasi-religious principle', and our task is now to 'desacralise' it, we need not worship at the altar of behavioural law instead.[8] In a democratic platform governance, the role of transparency should not only be to enforce the law, but also to consider carefully when and how the law is needed, alongside society's many non-legal forms of contestation and resistance.

8 David Heald and Christopher Hood and David Heald (eds.). In *Transparency: The key to better governance?* (Oxford University Press for the British Academy 2006). David Pozen, 'Transparency's Ideological Drift' (2018) 126 *Yale Law Journal* 100.

7

References

Literature

Aalberg T, Strömbäck J and De Vreese C, 'The framing of politics as strategy and game: A review of concepts, operationalizations and key findings' (2012) 13 *Journalism* 162.

Aggarwal C, *Recommender Systems: The textbook* (Springer 2016).

Ali M and others, 'Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes' (2019) 3 *Proceedings of the ACM on human-computer interaction* 1.

Albright J, 'Facebook and the 2018 Midterms: A Look at the Data – The Micro-Propaganda Machine', *Medium* (4 November 2018) <https://medium.com/s/the-micro-propaganda-machine/the-2018-facebook-midterms-part-i-recursive-ad-ccountability-ac090d276097> accessed 15 September 2022.

Albu O and Flyverblom M, 'Organizational Transparency: Conceptualizations, Conditions, and Consequences' (2019) 58 *Business & Society* 268.

Alloa E, 'Transparency: A magic concept of modernity' in Emmanuel Alloa and Dieter Thomä (eds.), *Transparency, society and subjectivity* (Palgrave Macmillan 2018).

Alloa E, 'Why transparency has little (if anything) to do with the age of enlightenment' in Emmanuel Alloa (ed.), *This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor* (Leuven University Press 2022).

Andringa P, 'Interactive: See Political Ads Targeted to You on Facebook', *NBC (16 January* 2018). <http://www.nbcsandiego.com/news/tech/New-Data-Reveal-Wide-Range-Political-Actors-Facebook-469600273.html> accessed 16 September 2022.

Angelopoulos C, *European Intermediary Liability in Copyright: A Tort-Based Analysis: A Tort-Based Analysis* (Kluwer 2016).

Angelopoulos C and others, 'Study of fundamental rights limitations for online enforcement through self-regulation' (Research Report Institute for Information Law 2015) <https://pure.uva.nl/ws/files/8763808/IVIR_Study_Online_enforcement_through_self_regulation.pdf> accessed 16 September 2022.

Angwin J, 'Make Algorithms Accountable', *The New York Times* (1 August 2016) <https://www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html> accessed 16 September 2022.

Appelman N, 'Algorithmic content moderation through an agonistic lens: contesting online exclusion', Paper presented at MANCEPT: Digital Democracy, Governance and Resistance in a Digital Era (7 September 2022.).

Appelman N, Quintais J and Fahy R, 'Using Terms and Conditions to apply Fundamental Rights to Content Moderation: Is Article 12 DSA a Paper Tiger?' (*Verfassungsblog* 1 September 2021) <https://verfassungsblog.de/power-dsa-dma-06/> accessed 16 September 2022.

Ananny M, 'Toward an ethics of algorithms: Convening, observation, probability, and timeliness' (2017) 41 *Science, Technology, & Human Values* 93.

Ananny M and Crawford K, 'A Liminal Press: Situating news app designers within a field of networked news production' (2015) 3 *Digital Journalism* 192.

Ananny M and Crawford K, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2018) 20 *New Media & Society* 973.

Are C, 'The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram' (2021) *Feminist Media Studies* 1 <https://doi.org/10.1080/14680777.2021.1928259> accessed 28 September 2022.

Ausloos J, 'GDPR Transparency as a Research Method' (2019) SSRN Working paper. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3465680> accessed 16 September 2022.

Ausloos J and Dewitte P, 'Shattering one-way mirrors – data subject access rights in practice' (2018) 8 *International Data Privacy Law* 4 < https://doi.org/10.1093/idpl/ipy001> accessed 28 September 2020.

Ausloos J, Leerssen P and Ten Thije P, 'Operationalizing Research Access in Platform Governance: What To Learn From Other Industries?' (Research Report AlgorithmWatch 2020) <https://algorithmwatch.org/en/governing-platforms-ivir-study-june-2020/#study> accessed 16 September 2020.

Baker EC, *Media concentration and democracy: Why ownership matters* (Cambridge University Press) Cambridge University Press.

Bakshy E, Messing S and Adamic L, 'Exposure to ideologically diverse news and opinion on Facebook' (2015) 348 *Science* 6239.

Bambauer D, 'Against Jawboning' (2015) 100 *Minnesota Law Review* 51.

Barendt E, *Freedom of speech* (Oxford University Press 2005).

Barrett B and Kreiss D, 'Platform transience: changes in Facebook's policies, procedures, and affordances in global electoral politics' 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1446> accessed 25 September 2022.

Barocas S, 'The Price of Precision: Voter Microtargeting and Its Potential Harms to the Democratic Process' (2012) *Proceedings of the First Edition Workshop on Politics, Elections and Data* 31.

Barwise P and Watkins L, 'The evolution of digital dominance: how and why we got to GAFA', in: Martin Moore and Damian Tambini (eds.), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (Oxford University Press 2018).

Baume S and Papadopoulos Y, 'Transparency: from Bentham's inventory of virtuous effects to contemporary evidence-based scepticism' (2018) 21 *Critical Review of International Social and Political Philosophy* 169.

Benkler Y, Faris R and Roberts H, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press 2018).

Benkler Y, 'A Free, Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate' (2011) 46 *Harvard Civil Rights-Civil Liberties Review* 311.

Beauchamp Z, 'Facebook blocked the spread of a liberal article because a conservative told it to', *Vox* (12 September 2018) <https://www.vox.com/policy-and-politics/2018/9/12/17848026/facebook-thinkprogress-weekly-standard> accessed 16 September 2022.

Bickert M, 'Combatting Vaccine Misinformation', Facebook Newsroom (7 March 2019). <https://newsroom.fb.com/news/2019/03/combatting-vaccine-misinformation/> accessed 16 September 2022.

Bickert M, 'Combatting Vaccine Misinformation', Facebook Newsroom (7 March 2019) <https://newsroom.fb.com/news/2019/03/combatting-vaccine-misinformation/> accessed 26 September 2022.

Biddle S, 'Revealed: Facebook's Secret Blacklist of 'Dangerous Individuals and Organizations', *The Intercept* (12 October 2021). <https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/> accessed 26 June 2022.

Bietti E, 'Consent as a Free Pass: Platform Power and the Limits of the Informational Turn' (2020) 40 *Pace Law Review* 307.

Bijker W, *Of Bicycles, Bakelites, and Bulbs* (MIT Press 1995).

Binns R and others, 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation' (2017) *International Conference on social informatics* 405.

Birkner T, Koenen E and Schwarzenegger C, 'A Century of Journalism History as Challenge Digital archives, sources, and methods' (2018) 6 *Digital Journalism* 1121.

Bishop S, 'Algorithmic Experts: Selling Algorithmic Lore on YouTube' (2020) 6 *New Media + Society* 1.

Bloch-Wehba H, 'Global platform governance: private power in the shadow of the state' (2019) 72 *SMU Law Review*27.

Bloch-Wehba H, 'Automation in moderation' (2020) 52 *Cornell International Law Journal* 41.

Bodó B and others, 'Tackling the Algorithmic Control Crisis: The Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents' (2018) 19 *Yale Journal of Law and Technology* 133.

Bodó B, Helberger N and De Vreese, 'Political micro-targeting: a Manchurian candidate or just a dark horse? Towards the next generation of political micro-targeting research' (2017) 6(4) *Internet Policy Review* <https://policyreview.info/articles/analysis/political-micro-targeting-manchurian-candidate-or-just-dark-horse> accessed 16 September 2022.

Boerman S, Kruikemeier S and Zuiderveen Borgesius F, 'Online Behavioral Advertising: A Literature Review and Research Agenda' (2017) 46 *Journal of Advertising* 363.

Bowers J, Sedenberg E and Zittrain J, 'Platform Accountability Through Digital 'Poison Cabinets'', Knight First Amendment Institute (13 April 2021). <https://cyber.harvard.edu/story/2021-04/platform-accountability-through-digital-poison-cabinets> accessed 16 September 2022.

Boyd D, 'Social network sites as networked publics: Affordances, dynamics, and implications', in: Zizi Papacharissi (ed.), *A Networked Self* (Routledge 2010).

Bossetta M, 'Scandalous design: How social media platforms' responses to scandal impacts campaigns and elections' (2020) 6(2) *Social Media+ Society* <https://doi.org/10.1177/2056305120924777> accessed 16 September 2022.

Bovens M, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 *European Law Journal* 447.

Breland A, 'AOC Asked Mark Zuckerberg About Facebook's Fact-Checking Process. He Didn't Give Her the Whole Truth' , *Mother Jones* (23 October 2019) <https://www.motherjones.com/politics/2019/10/aoc-zuckerberg-facebook-congress-daily-caller-fact-check-dodge/> accessed 15 September 2022.

Jack Brewster, 'Trump Campaign Facebook Ad Strategy: Paint Biden As A Socialist', *Forbes* (13 April 2020) < https://www.forbes.com/sites/jackbrewster/2020/04/13/trump-campaign-facebook-ad-strategy-paint-biden-as-a-socialist/?sh=1ea81f53322f> accessed 25 September 2022.

Barrett B and Kreiss D, 'Platform transience: changes in Facebook's policies, procedures, and affordances in global electoral politics' 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1446> accessed 25 September 2022.

Bridle J, 'Something is wrong on the internet', *Medium* (6 November 2017) <https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2> accessed 25 September 2022.

Bridy A, 'The Price of Closing the Value Gap: How the Music Industry Hacked EU Copyright Reform' (2020) 22 *Vanderbilt Journal of Entertainment and Technology Law* 323.

Bruns A, 'Produsage' (2007) *C&C '07: Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition* 99.

Brother Ali, 'Uncle Sam Goddamn' (2007) *The Undisputed Truth* [CD], Rhymesayers Entertainment.

Birnhack M and Elkin-Koren N. 'The Invisible Handshake: The Reemergence of the State in the Digital Environment' (2003) 8(6) *Virginia Journal of Law & Technology* <https://law.bepress.com/taulwps/art54/> accessed 25 September 2022.

Bruns A, *Are Filter Bubbles Real?* (Polity Press 2019).

Bruns A, 'After the "APIcalypse": Social Media Platforms and Their Fight against Critical Scholarly Research' (2019) 22 *Information, Communication & Society* 1544.

Bucher T, 'The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms' (2017) 20 *Information, communication & society* 30.

Burgess J, Marwick A and Poell T (eds.), *The SAGE Handbook of Social Media* (Sage 2017).

Burrell J, 'How the machine "thinks": Understanding opacity in machine learning algorithms' (2017) 3(1) *Big Data & Society* <https://doi.org/10.1177/2053951715622512> accessed 15 September 2022.

Caplan R, *Networked Platform Governance: Reconciling Horizontals and Hierarchies in the Platform Era* (Doctoral dissertation, Rutgers The State University of New Jersey, School of Graduate Studies 2021).

Caplan R and Gillespie T, 'Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy' (2020) 6(2) *Social Media+ Society* <https://doi.org/10.1177/2056305120936636> accessed 15 September 2022.

Carbajal A and others, 'Open Letter to Marck Zuckerberg on Alternative Solutions for Politics Tagging' (2018) News Media Alliance. <https://www.newsmediaalliance.org/wp-content/uploads/2018/06/vR_Alternative-Facebook-Politics-Tagging-Solutions-FINAL.pdf> accessed 15 September 2022.

Cardoso T, 'Google to ban political ads ahead of federal election, citing new transparency rules'. The Globe and Mail (March 4 2019). <https://www.theglobeandmail.com/politics/article-google-to-ban-political-ads-ahead-of-federal-election-citing-new/> accessed 15 September 2022.

Carlson A, 'The Need for Transparency in the Age of Predictive Sentencing Algorithms' (2017) 103 *Iowa Law Review* 303.

Carrero J, 'Access Granted: A First Amendment Theory of Reform of the CFAA Access Provisions' (2020) 120 *Columbia Law Review* 131.

Castendyk O, Dommering E and Scheuer A, *European Media Law* (Kluwer 2008).

Cellan-Jones R, 'Facebook's News Feed experiment panics publishers', *BBC News* (24 October 2017). <https://www.bbc.com/news/technology-41733119> accessed 15 September 2022.

Chadwick A, *The Hybrid Media System: Politics and power* (Oxford University Press 2017).

Chadwick A, Vaccari C and Kaiser J, 'The amplification of exaggerated and false news on social media: The roles of platform use, motivations, affect, and ideology' (2021) *American Behavioral Scientist* < https://doi.org/10.1177/00027642221118264> accessed 15 September 2022.

Chaslot G, 'YouTube's A.I. was divisive in the US presidential election', *Medium* (27 November 2016) <https://medium.com/the-graph/youtubes-ai-is-neutral-towards-clicks-but-is-biased- towards-people-and-ideas-3a2f643dea9a> accessed 15 September 2022.

Chavern D, 'Open Letter to Mr. Zuckerberg', *News Media Alliance* (18 May 2018). Retrieved from <http://www.newsmediaalliance.org/wp-content/uploads/2018/05/FB-Political-Ads-Letter-FINAL.pdf> accessed 15 September 2022.

Chester J and Montgomery K, 'The role of digital marketing in political campaigns' (2017) 6(4) *Internet Policy Review* <https://doi.org/10.14763/2017.4.773> accessed 15 September 2022.

Citron D and Pasquale F, 'The scored society: Due process for automated predictions' (2014) 89 *Washington Law Review* 1.

Cobbe J, 'Algorithmic censorship by social platforms: Power and resistance' (2021) 34 *Philosophy & Technology* 739.

Cobbe J and Singh J, 'Regulating Recommending: Motivations, Considerations, and Principles' (2019) 10(3) *European Journal of Law and Technology* https://ejlt.org/index.php/ejlt/article/view/686 accessed 15 September 2022.

Cohen J, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press 2019).

Cohen D, 'FBI Uses Facebook Ads in Washington, D.C., to Find Information on Russian Spies', *Adweek* (2 October 2019).

Coleman S and Ross K, *The Media and the Public* (Wiley 2010).

Collins B, 'Hillary PAC Spends $1 Million to 'Correct' Commenters on Reddit and Facebook', *The Daily Beast* (21 April 2016) <https://www.thedailybeast.com/articles/2016/04/21/hillary-pac-spends-1-million-to-correct-commenters-on-reddit-and-faceboook> accessed 15 September 2022.

Common M, 'Fear the reaper: How content moderation rules are enforced on social media' (2020) 34 *International Review of Law, Computers & Technology* 126.

Cornia A and others, *Private Sector News, Social Media Distribution ,and Algorithm Change* (Research Report Reuters Institute Digital News Project Report 2018) <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-10/Cornia_Private_Sector_News_FINAL.pdf> accessed 25 September 2020.

Cotter K, 'Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram' (2019) 21 *New Media & Society* 895.

Cotter, K, '"Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms' (2021) *Information, Communication & Society* <https://doi.org/10.1080/1369118X.2021.1994624> accessed 15 September 2022.

Crawford K, 'Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics' (2016) 41 *Science, Technology, & Human Values* 77.

Cucciniello M, Porombescu G and Grimmelikhuijsen S, '25 Years of Transparency Research: Evidence and Future Directions' (2017) 77 *Public Administration Review* 32.

De Gregorio G, 'Democratising online content moderation: A constitutional framework' (2020) 36 *Computer Law & Security Review* 105374.

De Vreese C and others, 'Public statement from the Co-Chairs and European Advisory Committee of Social Science One', Social Science One (11 December 2019) <https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one> accessed 15 September 2022.

Department for Digital, Culture, Media & Sport, Online Harms White Paper (2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf> accessed 27 September 2022.

Diakopoulos N, 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures' (2014) 3 *Digital Journalism* 398.

Diakopoulos N, 'Accountability in Algorithmic Decision Making' (2016) 59 *Communications of the ACM* 56.

Diakopoulos N, 'The Algorithms Beat: Angles and Methods for Investigation' in Jonathan Gray and Liliana Bounegru (eds.), *The Data Journalism Handbook 2.0* (Amsterdam University Press 2018).

Diakopoulos N, *Automating the News: How Algorithms are Rewriting the Media* (Harvard University Press 2019).

Diakopoulos N and Koliska M, 'Algorithmic Transparency in the News Media' (2016) 5 *Digital Journalism* 809.

Diresta R, 'Free Speech Is Not the Same As Free Reach', *Wired Magazine* (30 August 2018). https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/ accessed 28 June 2022.

Dobber T, Ó Fathaigh R and Zuiderveen Borgesius, 'The regulation of online political micro-targeting in Europe', 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1440> accessed 24 September 2022.

Domino J, 'Why Facebook's Oversight Board is Not Diverse Enough', *Just Security* (21 May 2020) <https://www.justsecurity.org/70301/why-facebooks-oversight-board-is-not-diverse-enough/ accessed 15 September 2022.

Dommett K, 'Regulating digital campaigning: The need for precision in calls for transparency' (2020) 12 *Policy & Internet* 432.

Dommett K, 'The inter-institutional impact of digital platform companies on democracy: A case study of the UK media's digital campaigning coverage' (2021) *New Media & Society* <https://doi.org/10.1177/14614448211028546> accessed 15 September 2022.

Donovan J and Boyd D, 'Stop the presses? Moving from strategic silence to strategic amplification in a networked media ecosystem' (2021) 65 *American Behavioral Scientist* 333.

Douek E, 'Content Moderation As Administration' (2022) 136 *Harvard Law Review*, forthcoming. < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4005326> accessed 15 September 2022.

Douek E, 'The limits of international law in content moderation' (2021) 6 *UC Irvine Journal of International, Transnational and Comparative Law* 37.

Dvoskin B, 'Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance' (2022) 67 *Villanova Law Review* 447.

Duffy BE, 'Algorithmic precarity in cultural work' (2020) 5 *Communication and the Public* 1.

Edelman B, 'Pitfalls and Fraud In Online Advertising Metrics: What Makes Advertisers Vulnerable to Cheaters, And How They Can Protect Themselves' (2014) 54 *Journal of Advertising Research* 127.

Edelson L and others, 'An Analysis of United States Online Political Advertising Transparency' (2019) *ArXiv [Cs]* <http://arxiv.org/abs/1902.04385> accessed 15 September 2022.

Edelson, L, Lauinger, T, & McCoy, D, 'A Security analysis of the Facebook Ad library' (2020) *IEEE Symposium on Security and Privacy* 661.

European Digital Media Observatory, Report of the Working Group on Platform-to-Researcher Data Access (2022) <https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf> accessed 26 September 2022.

Edwards L and Veale M, 'Slave to the algorithm? Why a "Right to an Explanation" is probably not the remedy you are looking for' (2017) 16 *Duke Law & Technology Review* 18.

Erdos D, 'Disclosure, Exposure and the "Right to be Forgotten" after Google Spain: Interrogating Google Search's webmaster, end user and Lumen notification practices' (2020) 38 *Computer Law & Security Review* 105437.

Eskens S, *The fundamental rights of news users: The legal groundwork for a personalised online news environment* (Doctoral Dissertation, University of Amsterdam, Faculty of Law 2021).

Etzioni A, 'Is Transparency the Best Disinfectant?' (2010) 18 *The Journal of Political Philosophy* 389.

Eslami M and others, 'First I "like" it, then I hide it: Folk Theories of Social Feeds' (2016) *CHI '16* 2371.

European Commission, Third monthly intermediate results of the EU Code of Practice against disinformation (2019) <https://ec.europa.eu/digital-single-market/en/news/third-monthly-intermediate-results-eu-code-practice-against-disinformation> accessed 27 September 2022.

European Commission, 'Commission launches call to create the European Digital Media Observatory' (2019) <https://digital-strategy.ec.europa.eu/en/news/commission-launches-call-create-european-digital-media-observatory> accessed 27 September 2022.

European Digital Media Observatory (EDMO), Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access (2022). <https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf> accessed 28 September 2022.

Ezrachi A and Stucke M, 'Virtual competition' (2016) 7 *Journal of European Competition Law & Practice* 585.

Facebook, 'How People Help Fight False News', *Facebook Newsroom* (21 June 2018) <https://newsroom.fb.com/news/2018/06/inside-feed-how-people-help-fight-false-news/> accessed 15 September 2022.

Facebook, 'About Social issues', *Facebook Business Help Center* (n.d.). Retrieved from <https://www.facebook.com/business/help/214754279118974> accessed 15 September 2022.

Facebook, 'Ads about social issues, elections or politics', *Facebook Business Help Center* (n.d.) <https://www.facebook.com/business/help/208949576550051> accessed 15 September 2022.

Facebook, 'Sharing Our Content Distribution Guidelines', *Facebook Newsroom* (23 September 2021) <https://perma.cc/BRT3-7XC8> accessed 28 June 2022.

Facebook, 'What is the Facebook Ad Library and how do I search it?', *Facebook Business Help* Center (n.d.) <https://www.facebook.com/business/help/214754279118974?id=288762101909005> accessed 15 September 2022.

Fair G and Wesslen R, 'Shouting into the void: A database of the alternative social media platform Gab' (2019) 13 *Proceedings of the International AAAI Conference on Web and Social Media* 608.

Fanta A and Rudl T, 'Leaked document: EU Commission mulls new law to regulate online platforms', *Netzpolitik* (16 July 2019) <https://netzpolitik.org/2019/leaked-document-eu-commission-mulls-new-law-to-regulate-online-platforms/#spendenleiste> accessed 15 September 2022.

Faris R and others, 'Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election' (Research Report Berkman Klein Center 2017) <https://dash.harvard.edu/bitstream/handle/1/33759251/2017-08_electionReport_0.pdf?sequence=9&isAllowed=y> accessed 15 September 2022.

Fertmann M and others, 'Hybrid institutions for disinformation governance: Between imaginative and imaginary', *Internet Policy Review* (16 May 2022) <https://policyreview.info/articles/news/hybrid-institutions-disinformation-governance-between-imaginative-and-imaginary/1669> accessed 25 September 2022.

Fisher M and Taub A, 'On YouTube's Digital Playground, an Open Gate for Pedophiles', *The New York Times* (3 June 2019) <https://www.nytimes.com/2019/06/03/world/americas/youtube- pedophiles.html> accessed 19 September 2022.

Goodman B and Flaxman S, 'European Union regulations on algorithmic decision-making and a "right to explanation"' (2017) 38 *AI Magazine* 3.

Flyverbom M, *The Digital Prism: Transparency and managed visibilities in a datafied world* (Cambridge University Press 2019).

Foucault M, *Discipline and Punish: The Birth of the Prison* (Pantheon Books 1977).

Fowler M and Brenner D, 'A Market Place Approach to Broadcast Regulation' (1982) 60 *Texas Law Review* 207.

François C and Douek E, 'The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations' (2021) 1 *Journal of Online Trust and Safety* 1.

Frier S, 'Facebook's Political Rule Blocks Ads for Bush's Beans, Singers Named Clinton', *Bloomberg* (2 July 2018) <https://www.bloomberg.com/news/articles/2018-07-02/facebook-s-algorithm-blocks-ads-for-bush-s-beans-singers-named-clinton> accessed 15 September 2022.

Frosio G (ed.), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

Fung A, Weil D and Graham M, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge University Press 2007).

Fung A, 'Infotopia: Unleashing the Democratic power of transparency' (2013) 41 *Politics & Society* 183.

Fulgoni G, 'Fraud in Digital Advertising: A Multibillion-Dollar Black Hole: How Marketers Can Minimize Losses Caused by Bogus Web Traffic' (2016) 56 *Journal of Advertising Research* 122.

Freelon D, 'Computational research in the post-API age' (2018) 35 *Political Communication* 4.

Secretary of State for Digital Affairs of the Republic of France, 'Creating a French framework to make social media platforms more accountable: Final Mission Report on the Regulation of social networks (2019) <https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf> accessed 26 September 2022.

Garton Ash T, *Free Speech: Ten Principles for a Connected World* (Yale University Press 2016).

Gary J and Soltani A, 'First Things First: Online Advertising Practices and Their Effects on Platform Speech', *Knight First Amendment Institute* (21 August 2019) <https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech> .

Ghosh A, Venkatadri G. and Mislove A, 'Analyzing Political Advertisers' Use of Facebook's Targeting Features' (2019) *IEEE workshop on technology and consumer protection* <https://www.ieee-security.org/TC/SPW2019/ConPro/papers/ghosh-conpro19.pdf>

Giglietto F and others, 'It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections' (2020) 23 *Information, Communication & Society* 867.

Gillespie T, 'The relevance of algorithms', in Tarleton Gillespie and others (eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (MIT Press 2014).

Gillespie T, 'Platforms Intervene' (2015) 1 *Social Media + Society* 1.

Gillespie T, 'Governance of and by platforms', in Jean Burgess, Alice Marwick and Thomas Poell (eds), *The SAGE handbook of social media* (Sage 2017).

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

Gillespie T, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 *Social Media+ Society* <https://doi.org/10.1177/205630512211175> accessed 19 September 2022.

Gitlin T, 'Public sphere or public sphericules?' in James Curran and Tamar Liebes (eds), *Media Ritual and Identity* (Routledge 1998)

Glasser T, 'Three views on accountability', in Everette Dennis, Donald Gillmor and Theodore Glasser (eds.), *Media Freedom and Accountability* (Praeger 1989).

Goldman R, 'Update on Our Advertising Transparency and Authenticity Efforts', Facebook Newsroom (27 October 2017) <https://newsroom.fb.com/news/2017/10/update-on-our-advertising-transparency-and-authenticity-efforts/> accessed 19 September 2022.

Google, Implementation Report for EU Code of Practice on Disinformation (2019) <https://ec.europa.eu/information_society/newsroom/image/document/2019-5/google_-_ec_action_plan_reporting_CF1r62236-E8FB-725E-C0A3D2D6CCFE678A_56994.pdf> accessed 26 September 2022.

Google, 'Verification for election advertising in the European Union' (n.d.) <https://support.google.com/adspolicy/answer/9211218> accessed 26 September 2022.

Goldman E, 'Content Moderation Remedies' (2021) 28 *Michigan Technology Law Review* 1.

Goldmacher S, 'Biden Pours Millions Into Facebook Ads, Blowing Past Trump's Record', *The New York Times* (2 September 2020).

Gorwa R, 'What is platform governance?' (2019) 22 *Information, Communication & Society* 6.

Gorwa R, 'The platform governance triangle: conceptualising the informal regulation of online content' (2019) 8(2) *Internet Policy Review* 2 <https://doi.org/10.14763/2019.2.1407> accessed 19 September 2022.

Gorwa R, 'Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG' (2021) 56 *Telecommunications Policy* 102145.

Gorwa R and Garton Ash T, 'Democratic Transparency in the Platform Society' in: Nate Persily and Joshua Tucker (eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge University Press 2020).

Gorwa R, Binns R and Katzenbach C, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) *Big Data & Society* 1 <https://doi.org/10.1177/2053951719897945> accessed 19 September 2022.

Gorton W, 'Manipulating citizens: How political campaigns' use of behavioral social science harms democracy' (2020) 38 *New Political Science* 61.

Green L and Adams T, 'Legal Positivism', *The Stanford Encyclopedia of Philosophy* (2019) <https://plato.stanford.edu/archives/win2019/entries/legal-positivism/> accessed 25 September 2022.

Griffin A, 'Tiktok user reveals ingenious Facebook trick to find hidden discount codes', *The Independent* (2020, August 11) <https://www.independent.co.uk/life-style/gadgets-and-tech/news/discount-codes-facebook-ad-library-tik-tok-a9665221.html> accessed 19 September 2022.

Griffin R, 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality', SSRN Draft Paper (2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064738> accessed 19 September 2022.

Griffin R, 'The Sanitised Platform' (2022) 13 *JIPITEC* 36.

Grimmelmann J, 'The virtues of moderation' (2015) 17 *Yale Journal of Law and Technology* 42.

Grimmelman J & Westreich W, 'Incomprehensible Discrimination', 7 *California Law Review Online* https://scholarship.law.cornell.edu/facpub/1536/> accessed 12 September 2022.

Grygiel J and Sager W, 'Unmasking Uncle Sam: A Legal test for identifying State media' (2020) 11 *UC Irvine Law Review* 383.

Guha S, Cheng B and Francis P, 'Challenges in measuring online advertising systems' (2020). *Proceedings of the 10th Annual Conference on Internet Measurement - IMC '10* 81.

Hansen H, Christensen L and Flyverbom, 'Logics of transparency in late modernity: paradoxes, mediation and governance' (2015) 18 *European Journal of Social Theory* 117.

Hao K, 'YouTube is experimenting with ways to make its algorithm even more addictive', *MIT Technology Review* (27 September 2019) <https://www.technologyreview.com/s/614432/youtube-algorithm-gets-more-addictive/> accessed 17 September 2022.

Harambam J, Helberger N and Van Hoboken J, 'Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem' (2018) 376 *Philosophical Transactions of the Royal Society* 2133.

Hargreaves E and others, 'Biases in the Facebook News Feed: a Case Study on the Italian Elections' (2018) *International Symposium on Foundations of Open Source Intelligence and Security Informatics, In conjunction with IEEE/ACM ASONAM* <https://hal.inria.fr/hal-01907069> accessed 19 September 2022.

Hargreaves E and others, 'Fairness in Online Social Network Timelines: Measurements, Models and Mechanism Design', 127 *Performance Evaluation Review* 15.

Hartzog W and Stutzman F, 'The Case for Online Obscurity' (2013) 101 *California Law Review* 1.

Heald D, 'Varieties of transparency' in David Heald and Christopher Hood and David Heald (eds.) *Transparency: The key to better governance?* (Oxford University Press for the British Academy 2006).

Heawood J, 'Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal' (2018) 23 *Information Polity* 429.

Helberger N, 'Diversity by Design' (2011) 1 *Journal of Information Policy* 441.

Helberger N, 'Exposure diversity as a policy goal' (2012) 4 *Journal of Media Law* 65.

Helberger N, 'On the democratic role of news recommenders' (2019) 7 *Digital Journalism* 993.

Helberger N, 'The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power' (2022) 8 *Digital Journalism* 842.

Helberger N, Karppinen K and d'Acunto L, 'Exposure diversity as a design principle for recommender systems' (2018) 21 *Information, Communication & Society* 191.

Helberger N, Kleinen-Von Königslöw K and Van der Noll R, 'Regulating the new information intermediaries as gatekeepers of information diversity' (2015) 17 *info* 50.

Helberger N, Pierson J and Poell T, 'Governing online platforms: From contested to cooperative responsibility' (2018) 34 *The Information Society* 1.

Heldt A, 'Borderline speech: caught in a free speech limbo?', *Internet Policy Review* (15 October 2021) <https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510> accessed 15 September 2020.

Helmond A and Van der Vlist F, 'Social media and platform historiography: Challenges and opportunities' (2019) 22 *TMG–Journal for Media History* 6.

Helmond A, 'The platformization of the web: Making web data platform ready' (2015) 1(2) *Social Media + Society* <https://doi.org/10.1177/2056305115603080> accessed 19 September 2022.

Hern A, 'Facebook to block foreign spending on Irish abortion vote ads', *The Guardian* (8 May 2018) <https://www.theguardian.com/world/2018/may/08/facebook-to-block-foreign-spending-on-irish-abortion-vote-ads-referendum> accessed 24 September 2022.

Hofmann J, Katzenbach C and Gollatz K, 'Between coordination and regulation: Finding the governance in Internet governance' (2017) 19 *New Media & Society* 1406.

Hood C and Heald D (eds.), *Transparency: The key to better governance?* (Oxford University Press for the British Academy 2006).

Horten M, 'Algorithms Patrolling Content: Where's the Harm? An empirical examination of Facebook shadow bans and their impact on users' (2021) SSRN Draft Paper <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792097> accessed 28 June 2022.

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/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353> accessed 28 June 2022.

Hounsel A and others, 'Estimating Publication Rates of Non-Election Ads by Facebook and Google', *Github* (1 November 2019). <https://github.com/citp/mistaken-ad-enforcement/blob/master/estimating-publication-rates-of-non-election-ads.pdf> accessed 19 September 2022.

House of Commons Select Committee on Digital, Culture, Media and Sport, Disinformation and 'Fake News' (Final Report 2019) <https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179103.htm#_idTextAnchor000> accessed 22 September 2022.

Howard P and others, 'The IRA, Social Media and Political Polarization in the United States, 2012–2018' (Research Report Oxford Computational Propaganda Project 2018) <https://www.oii.ox.ac.uk/news-events/reports/the-ira-social-media-and-political-polarization-in-the-united-states-2012-2018/> accessed 15 September 2022.

Howard P, 'A Way to Detect the Next Russian Misinformation Campaign', *The New York Times* (27 March 2019) <https://www.nytimes.com/2019/03/27/opinion/russia-elections-facebook.html?module=inline> accessed 19 September 2022.

Husovec M, *Injunctions against intermediaries in the European Union: Accountable but not liable?* (Cambridge University Press 2017).

Husovec M and Roche Laguna I, 'Digital Services Act: A Short Primer', in: Husovec and Roche Laguna, *Principles of the Digital Services Act* (Oxford University Press forthcoming 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4153796> accessed 25 September 2022.

InfluenceMap, 'Big Oil's Real Agenda on Climate Change (Research Report InfluenceMap 2019) <https://influencemap.org/report/How-Big-Oil-Continues-to-Oppose-the-Paris-Agreement-38212275958aa21196dae3b76220bddc> accessed 19 September 2022.

Jaidka K, Mukerjee S and Lelkes Y, 'Censorship on social media: The gatekeeping functions of shadowbans in the American Twitterverse' (2022) SSRN Draft Paper <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4087843> accessed 28 June 2022.

Kaminski M, 'The Right to Explanation, Explained' (2019) 34 *Berkeley Technology Law Journal* 189.

Kaminski M, 'Understanding Transparency in Algorithmic Accountability' in Woodrow Barfield (ed.), *Cambridge Handbook of the Law of Algorithms* (Cambridge University Press 2020).

Karppinen K, *Rethinking Media Pluralism* (Fordham University Press 2013).

Keller D, 'Who Do You Sue? State and Platform Hybrid Power over Speech' (2019) Hoover Institution Aegis Series 1902. <https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf> accessed 15 September 2022.

Keller D, 'Amplification and its discontents: why regulating the reach of online content is hard', *Knight First Amendment Institute* (8 June 2021) <https://knightcolumbia.org/content/amplification-and-its-discontents> accessed 25 September 2022.

Keller D and Leerssen P, 'Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation', in: Persily N. and Tucker J (eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge University Press 2020).

Kemper J and Kolkman D, 'Transparent to whom? No algorithmic accountability without a critical audience' (2019) 22 *Information, Communication & Society* 2081.

Kettemann M and Tiedeke A, 'Back up: Can users sue platforms to reinstate deleted content?' (2020) 9(2) *Internet Policy Review* <https://doi.org/10.14763/2020.2.1484> accessed 19 September 2022.

Kim Y and others, 'The stealth media? Groups and targets behind divisive issue campaigns on facebook'(2018) 35 *Political Communication* 515.

Kleis Nielsen R and Ganter S, *The Power of Platforms: Shaping Media and Society* (Oxford University Press 2022).

Klimkiewicz B, 'Is the Clash of Rationalities Leading Nowhere? Media Pluralism in European Regulatory Policies', in Andrea Czepek, Melanie Hellwig and Eva Nowak (eds.), *Press Freedom and Pluralism in Europe: Concepts and Conditions* (University of Chicago Press 2009).

Klonick K, 'The New Governors: The people, rules, and processes governing online speech' (2017) 131 *Harvard Law Review* 1598.

Koivisto I, *The Transparency Paradox: Questioning an ideal* (Oxford University Press 2022).

Korkea-aho E and Leino P, 'Who owns the information held by EU agencies? Weed killers, commercially sensitive information and transparent and participatory governance' (2017) 57 *Common Market Law Review* 1059.

Krafft T, Gamer M and Zweig K, 'Wer sieht was? Personalisierung, Regionalisierung un die Frage nach der Filterblase in Googles Suchmacschine' (Research Report Project #Datenspende 2018) <https://www.blm.de/files/pdf2/bericht-datenspende---wer-sieht-was-auf-google.pdf> accessed 19 September 2022.

Krebs LM and others, 'Tell Me What You Know: GDPR Implications on Designing Transparency and Accountability for News Recommender Systems' (2019) *CHI EA '19* < https://doi.org/10.1145/3290607.3312808> accessed 25 September 2022.

Kreimer S, 'The freedom of information act and the ecology of transparency' (2007) 10 *University of Pennsylvania Journal of Constitutional Law* 1011.

Kreiss D and Barrett B, 'Democratic tradeoffs: Platforms and political advertising' (2020) 16 *Ohio State Technology Law Journal* 493.

Kusche I, 'Private Voting, Public Opinion and Political Uncertainty in the Age of Social Media' (2022) 51 *Zeitschrift für Soziologie* 83.

Kuczerawy A, 'Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?', in: Elzbieta Kużelewska and others (eds.), *Disinformation and Digital Media as a Challenge for Democracy* (Intersentia 2021).

Kwoka M, 'FOIA, Inc.' (2016) 65 *Duke Law Journal* 1361.

Lada A, Wang M and Yan T, 'How Does News Feed Predict What You Want to See?', *Meta Newsroom* (26 January 2021) <https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/> accessed 28 June 2022.

Laidlaw E, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press 2015).

Latour B, *Science in Action* (Harvard University Press 1987).

Lapowsky I, 'Obscure Concealed-Carry Group Spent Millions on Facebook Political Ads', *WIRED* (19 January 2018) <https://www.wired.com/story/facebook-ads-political-concealed-online/> accessed 19 September 2022.

Le Merrer E, Morgan B and Trédan G, 'Setting the record straighter on shadow banning' (2021) *IEEE INFOCOM 2021-IEEE Conference on Computer Communications* 1.

Leathern R, 'Updates to our ad transparency and authorisation efforts' (29 November 2018). <https://www.facebook.com/facebookmedia/blog/updates-to-our-ads-transparency-and-authorisation-efforts> accessed 15 September 2019.

Ledwich M and Zaitsev A, 'Algorithmic extremism: Examining YouTube's rabbit hole of radicalization' (2019) 25(3) *First Monday* <https://doi.org/10.5210/fm.v25i3.10419 > accessed 19 September 2022.

Leerssen P, 'Cut out by the middle man: the free speech implications of social media blocking and banning in the EU' (2015) 6 *JIPITEC* 99.

Leerssen P, 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems', 11(2) *EJLT* < https://ejlt.org/index.php/ejlt/article/view/786> accessed 24 September 2020.

Leerssen P, 'Platform ad archives in Article 30 DSA', *DSA Observatory Blog* (25 May 2021) <https://dsa-observatory.eu/2021/05/25/platform-ad-archives-in-article-30-dsa/>.

Leerssen P and others, 'Platform ad archives: promises and pitfalls' (2019) 8(4) *Internet Policy Review* <https://doi.org/10.14763/2019.4.1421> accessed 5 November 2022.

Leerssen P and others, 'News from the ad archive: How journalists use the Facebook Ad Lbirary hold online advertising accountable'. *Information Communication & Society* < https://doi.org/10.1080/1369118X.2021.2009002>.

Leiser M, 'AstroTurfing, 'CyberTurfing' and other online persuasion campaigns' (2016) 7(1) *European Journal of Law and Technology* <https://ejlt.org/index.php/ejlt/article/view/501> accessed 19 September 2022.

Levi-Faur D, 'Regulation and Regulatory Governance', in: David Levi-Faur (ed.), *Handbook on the Politics of Regulation* (Edward Elgar 2011).

Lewis P, 'Fiction is outperforming reality': how YouTube's algorithm distorts truth', *The Guardian* (2 February 2018) <https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth> accessed 19 September 2022.

Lewis R, 'Alternative Influence: Broadcasting the Reactionary Right on YouTube' (Data & Society Research Report 2018) <https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf > accessed 19 September 2022.

Lewis R, 'All of YouTube, Not Just the Algorithm, is a Far-Right Propaganda Machine', *FFWD* (8 January 2020) <https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right- propaganda-machine-29b07b12430> accessed 22 September 2020.

Lindstedt C and Naurin D, 'Transparency is Not Enough: Making Transparency Effective in Reducing Corruption' (2010) 31 *International Political Science Review* 301.

Livingstone S and others, Tackling the Information Crisis: A Policy Framework for Media System Resilience (Report of the LSE Commission on Truth, Trust and Technology 2018) <http://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf> accessed 19 September 2020.

Lomas N, 'Facebook finally hands over leave campaign Brexit ads', *Techcrunch* (26 July 2018) <https://techcrunch.com/2018/07/26/facebook-finally-hands-over-leave-campaign-brexit-ads/> accessed 19 September 2022.

Lourenco RP, 'Evidence of an open government data portal impact on the public sphere' (2016) 12 *International Journal of Electronic Government Research* 21.

Lulamae J, 'How the "Shadow Banning" Mystery is Messing with Climate Activists' Heads', *AlgorithmWatch* (11 February 2022) <https://perma.cc/JYC5-BQ82> accessed 28 June 2022.

Lumb D, 'Why scientists are upset about the Facebook Filter Bubble story', *Fast Company* (5 August 2015) <http://www.fastcompany.com/3046111/fast-feed/why-scientists-are-upset-over-the-facebook-filter-bubble-study> accessed 19 September 2022.

Lyon D, *Theorizing Surveillance: The Panopticon and Beyond* (Willan Publishing 2006).

Macleod A, 'Fake News, Russian Bots and Putin's Puppets', in Alan MacLeod (ed.), *Propaganda in the Information Age: Still Manufacturing Consent* (Routledge 2019).

Mahieu R and Ausloos J, 'Harnessing the collective potential of GDPR access rights: towards an ecology of transparency', *Internet Policy Review* (6 July 2020) <https://policyreview.info/articles/news/harnessing-collective-potential-gdpr-access-rights-towards-ecology-transparency/1487> accessed 22 September 2022.

Mares R, 'Corporate transparency laws: A hollow victory?' (2018) 36 *Netherlands Quarterly of Human Rights* 189.

Marsden C, *Internet co-regulation: European law, regulatory governance and legitimacy in cyberspace* (Cambridge University Press 2011).

Marsden C, Meyer T and Brown I, 'Platform values and democratic elections: How can the law regulate digital disinformation?' (2019) 36 *Computer Law & Security Review* 105373.

Mashaw J, 'Accountability and Institutional Design: Some Thoughts on the Grammar of Governance' (2006) Yale Law School Public Law Working Paper No. 116

Matias J, Hounsel A and Hopkins, 'We Tested Facebook's Ad Screeners and Some Were Too Strict', *The Atlantic* (2 November 2018) <https://www.theatlantic.com/technology/archive/2018/11/do-big-social-media-platforms-have-effective-ad-policies/574609/> accessed 19 September 2022.

May P, 'Regulatory regimes and accountability' (2007) 1 *Regulation & Governance* 8.

Mayring P, 'Qualitative content analysis' (2000) 1 *Forum: Qualitative Social Research* 159.

Mazzucato M, 'Let's Make Data Into A Public Good', *MIT Policy Review* (27 June 2018). <https://www.technologyreview.com/s/611489/lets-make-private-data-into-a-public-good/> accessed 22 November 2022.

McCormick R, 'This image of Mark Zuckerberg says so much about our future', *The Verge* (22 February 2022) <https://www.theverge.com/2016/2/22/11087890/mark-zuckerberg-mwc-picture-future-samsung> accessed 28 September 2022.

McCracken G, *The Long Interview* (Sage 1988).

McGonagle T, 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation', in Giancarlo Frosio (ed.), *The Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

Meijer A, 'Transparency', in Mark Bovens, Robert Goodin and Thomas Schillemans (eds.), *The Oxford Handbook of Public Accountability* (Oxford University Press 2014).

Merrick R, 'Brexit: Leave "very likely" won EU referendum due to illegal overspending, says Oxford professor's evidence to High Court', *The Independent* (25 December 2019) <https://www.independent.co.uk/news/uk/politics/vote-leave-referendum-overspending-high-court-brexit-legal-challenge-void-oxford-professor-a8668771.html> accessed 22 September 2022.

Merrill J, 'How Big Oil Dodges Facebook's New Ad Transparency Rules', *ProPublica* (1 November 2018) <https://www.propublica.org/article/how-big-oil-dodges-facebooks-new-ad-transparency-rules> accessed 19 September 2022.

Merrill J and Tobin A, 'Facebook Moves to Block Ad Transparency Tools —Including Ours', *ProPublica* (28 January 2019) <https://www.propublica.org/article/facebook-blocks-ad-transparency-tools> accessed 19 September 2022.

Merrill J and Oremus W, 'Five points for anger, one for a "like": How Facebook's formula fostered rage and misinformation'. *The Washington Post* (26 October 2021) <https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/> accessed 19 September 2022.

Meta, Q4 2017 Earnings Report, *Meta Investor Relations* (31 January 2018) <https://investor.fb.com/investor-events/event-details/2018/Facebook-Q4-2017-Earnings/default.aspx> accessed 26 September 2022.

Michener G, 'Gauging the Impact of Transparency Policies' (2019) 79 *Public Administration Review* 136.

Miles C, 'What data is Crowdtangle tracking?', *CrowdTangle* (2022) <https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking> accessed 26 September 2022.

Miller J, 'Regulating robocalls: Are automated calls the sound of, or a threat to, democracy?' (2009) 16 *Michigan Technology Law Review* 213.

Mittelstadt B, 'Automation, algorithms, and politics: auditing for transparency in content personalization systems' (2016) 10 *International Journal of Communication* 12.

Montellaro Z, 'House Democrats forge ahead on electoral reform bill', *Politico* (26 February 2019) <https://politi.co/2GO4eJ8> accessed 19 September 2022.

Moore M and Tambini D, *Digital Dominance: The power of Google, Amazon, Facebook, and Apple* (Oxford University Press 2018).

Morozov E, *To Save Everything, Click Here* (Public Affairs 2014).

Morozov E, 'Digital Socialism? The Calculation Debate in the Age of Big Data' (2019) 116 *New Left Review*.

Mosseri A, 'Bringing People Close Together', Facebook Newsroom (11 January 2018) <https://newsroom.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/> accessed 19 September 2022.

Mozilla, 'Facebook and Google: This is What an Effective Ad Archive API Looks Like', *The Mozilla Blog* (2019, March 27) <https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like> accessed 19 September 2022.

Mozilla, 'Data Collection Log — EU Ad Transparency Report' (Research Report Mozilla 2019) <https://adtransparency.mozilla.org/eu/log/> accessed 19 September 2022.

Mulgan R, 'Accountability: An Ever-Expanding Concept?' (2000) 78 *Public Administration* 555.

Mulgan R, 'Comparing Accountability in the Public and Private Sectors' (2000) 59 *Australian Journal of Public Administration* 87.

Munger K and Philips J, 'Right-Wing YouTube: A Supply and Demand Perspective' (2020) 27 *The International Journal of Press/Politics* 186.

Myers West S, 'Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms' (2018) 20 *New Media & Society* 4366.

Napoli P, *Foundations of communications policy: Principles and process in the regulation of electronic media* (Hampton Press 2001).

Napoli P, 'Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers' (2015) 39 *Telecommunications Policy* 751.

Napoli P, *Social media and the public interest: Media regulation in the disinformation age* (2019 Columbia University Press).

Napoli P and Caplan R, 'Why media companies insist they're not media companies, why they're wrong, and why it matters' (2017) 22(5) *First Monday* <https://doi.org/10.5210/fm.v22i5.7051> accessed 19 September 2022.

Natale S, 'Amazon can read your mind: A media archaeology of the algorithmic imaginary', in Simone Natale and Diana Pasulka (eds.), *Believing in Bits: Digital Media and the Supernatural* (Oxford University Press).

Nicholas G, 'Shadowbanning Is Big Tech's Big Problem', *The Atlantic* (28 April 2022) <https://www.theatlantic.com/technology/archive/2022/04/social-media-shadowbans-tiktok-twitter/629702/> accessed 19 September 2022.

Noble S, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press 2018).

Norris P, 'Watchdog journalism', in Mark Bovens, Robert Goodin and Thomas Schillemans (eds.), *The Oxford Handbook of Public accountability* (Oxford University Press 2014).

Novak M, 'Trump's New Facebook Ads Claim He's Peacenik Who Also Loves Assassinations', *Gizmodo* (7 August 2020) <https://gizmodo.com/trumps-new-facebook-ads-claim-hes-peacenik-who-also-lov-1844644446> accessed 27 September 2022.

Nover S, 'Facebook Removes Trump Campaign Ads for Including Symbol Used by Nazis', *Adweek* (18 June 2020) <https://www.adweek.com/programmatic/facebook-removes-trump-campaign-ads-symbol-nazis/> accessed 26 September 2022.

Nunez M, 'Former Facebook Workers: We Routinely Suppressed Conservative News', *Gizmodo* (5 September 2016) <https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006> accessed 19 September 2022.

Ó Fathaigh R, Helberger N and Appelman N, 'The perils of legally defining disinformation' (2021). 10 *Internet Policy Review* <https://doi.org/10.14763/2021.4.1584> accessed 19 September 2022.

O'Callaghan D and others, 'Down the (white) rabbit hole: The extreme right and online recommender systems' (2015) 33 *Social Science Computer Review* 459.

O'Sullivan D, 'What an anti-Ted Cruz meme page says about Facebook's political ad policy', *CNN* (25 October 2018) <https://www.cnn.com/2018/10/25/tech/facebook-ted-cruz-memes/index.html> accessed 19 September 2022.

Ohm P (2010), 'Broken Promises of Privacy: Responding To The Surprising Failure of Anonymization' (2010) 57 *UCLA Law Review* 1701.

Oster J, *Media Freedom as a Fundamental Right* (Cambridge University Press 2015).

Pagnamenta R, 'Facebook will rue its left-wing oversight board appointments', *The Telegraph* (6 May 2020) <https://www.telegraph.co.uk/technology/2020/05/06/facebook-will-rue-left-wing-oversight-board-appointments/> accessed 19 September 2022.

Pariser E, *The Filter Bubble: What the internet is hiding from you* (Penguin 2011).

Parsons C, 'The (In)effectiveness of Voluntarily Produced Transparency Reports' (2019) 58 *Business & Society* 103.

Pasquale F, 'Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power' (2016) 17 *Theoretical Inquiries in Law* 487.

Pasquale F, *The Black Box Society: The secret algorithms that control money and information* (Harvard University Press 2015).

Peguera M, 'The DMCA Safe Harbors and Their European Counterparts: a comparative analysis of some common problems' (2009) 32 *Columbia Journal of Law & the Arts* 481.

Petre C, Duffy B and Hund E, '"Gaming the system": Platform paternalism and the politics of algorithmic visibility' (2019) 5(4) *Social Media+ Society* <https://doi.org/10.1177/20563051198799> accessed 25 September 2022.

Poell T, Nieborg, D and Duffy B, *Platforms and cultural production* (John Wiley & Sons 2021).

Poell T, Rajagopalan S and Kavada A, 'Publicness on platforms: Tracing the mutual articulation of platform architectures and user practices', in Zizi Papacharissi (ed.), *A Networked Self and Platforms, Stories, Connections* (Routledge 2018).

Poell T and Van Dijck J, 'Social Media and New Protest Movements', in: Jean Burgess, Alice Marwick and Thomas Poell (eds.), *The SAGE Handbook of Social Media* (Sage 2017).

Pozen D, 'Transparency's Ideological Drift' (2018) 126 *Yale Law Journal* 100.

Quarati A and De Martino M, 'Open government data usage: a brief overview' (2019) *IDEAS '19: Proceedings of the 23rd International Database Applications & Engineering Symposium*.

Quintais J, 'The new copyright in the Digital Single Market Directive: a critical look' (2020) 42 *European Intellectual Property Review* 28.

Radsch C, 'Shadowban / Shadow-ban', in: Luca Belli, Nicolo Zingales and Yasmin Curzi (eds), *IGF Glossary of Platform Law and Policy Terms* (Internet Governance Forum 2022) <https://platformglossary.info/> accessed 19 September 2022.

Remkes J and others, Lage Drempels, Hoge Dijken: Eindrapport (Staatscommissie Parlementair Stelsel 2018) <staatscommissieparlementairstelsel.nl/documenten/rapporten/samenvattingen/12/13/eindrapport> accessed 26 September 2022.

Rezende I, 'Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective' (2020) 11 *New Journal of European Criminal Law* 375.

Ricci F, Rokach L and Shapira B (eds.), *Recommender Systems Handbook* (Springer 2015).

Rieder B, Matamoroz-Fernandez A and Coromina O, 'From ranking algorithms to "ranking cultures": Investigating the modulation of visibility in YouTube search results' (2018) 24 *Convergence: The International Journal of Research into New Media Technologies* 50.

Rieder B, Coromina Ò and Matamoros-Fernández A, 'Mapping YouTube: A quantitative exploration of a platformed media system' (2020) 25 *First Monday* 8.

Rieder B and Hoffman J, 'Towards Platform Observability' (2020) 9(4) *Internet Policy Review* <https://doi.org/10.14763/2020.4.1535> accessed 19 September 2022.

Rieke A and Bogen M, Leveling the Platform: Real Transparency for Paid Messages on Facebook. (UpTurn Research Report 2018) <https://www.upturn.org/static/reports/2018/facebook-ads/files/Upturn-Facebook-Ads-2018-05-08.pdf> accessed 19 September 2022.

Riffe D and others, *Analyzing media messages. Using quantitative content analysis in research* (Routledge 2014).

Roberts S, *Behind the Screen: Content moderation in the shadows of social media* (Yale University Press 2021).

Rogers R, 'Social media research after the fake news debacle' (2018) 11 *Partecipazione e conflitto* 557.

Rogers R and Niederer S, *The Politics of Social Media Manipulation* (Amsterdam University Press 2020).

Roose K, 'In Virginia House Race, Anonymous Attack Ads Pop Up on Facebook', *The New York Times* (17 October 2018) <https://www.nytimes.com/2018/10/17/us/politics/virginia-race-comstock-wexton-facebook-attack-ads.html> accessed 26 September 2022.

Roose K, 'Inside Facebook's Data Wars', *The New York Times* (14 July 2021) <https://www.nytimes.com/2021/07/14/technology/facebook-data.html> accessed 6 November 2022.

Rosen G, 'Remove, Reduce, Inform: New Steps to Manage Problematic Content', *Facebook Newsroom* (2019) <https://newsroom.fb.com/news/2019/04/remove-reduce-inform-new-steps/> accessed 15 September 2022.

Rosenberg M, 'Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised', *The New York Times* (25 July 2019) <https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html> accessed 19 September 2022.

Rupar A, 'Facebook's controversial fact-checking partnership with a Daily Caller-funded website, explained', *Vox* (2 May 2019) <https://www.vox.com/2019/5/2/18522758/facebook-fact-checking-partnership-daily-caller> accessed 10 September 2022.

Rushkoff D, *Media virus! hidden agendas in popular culture* (Random House 1996).

Safarov I, Meijer A and Grimmelikhuijsen S, 'Utilization of open government data: A systematic literature review of types, conditions, effects and users' (2017) 22 *Information Polity* 1.

Sander B, 'Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation' (2019) 43 *Fordham International Law Journal* 939.

Sanders E, 'Washington Public Disclosure Commission Passes Emergency Rule Clarifying That Facebook and Google Must Turn Over Political Ad Data', *The Stranger* (9 May 2019) <https://www.thestranger.com/slog/2018/05/09/26158462/washington-public-disclosure-commission-passes-emergency-rule-clarifying-that-facebook-and-google-must-turn-over-political-ad-data> accessed 19 September 2022.

Sanders E, 'Facebook Says It's Immune from Washington State Law', *The Stranger* (16 October 2018) <https://www.thestranger.com/slog/2018/10/16/33926412/facebook-says-its-immune-from-washington-state-law> accessed 19 September 2022.

Sandvig C and others, 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms' (2014), Paper presented to *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* <https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf> accessed 26 September 2022.

Sax M, 'Algorithmic News Diversity and Democratic Theory: Adding Agonism to the Mix' (2022) *Digital Journalism* <https://doi.org/10.1080/21670811.2022.2114919> accessed 25 September 2022.

Schlesinger P, 'After the post-public sphere' (2014) 42 *Media, Culture & Society* 1545.

Scholz T, *Platform Cooperativism: Challenging the Corporate Sharing Economy* (Rosa Luxemburg Stiftung 2016).

Schulz W, Held T and Laudien A, 'Search Engines as Gatekeepers of Public Communication: Analysis of the German framework applicable to Internet search engines including media law and anti-trust law' (2005) 6 *German Law Journal* 1419.

Schulz W and Dreyer S, Governance von Informations-Intermediären - Herausforderungen und Lösungsansätze - Bericht an das BAKOM (Hans Bredow Institut Research Report 2020) <https://leibniz-hbi.de/de/publikationen/governance-von-informations-intermediaeren-herausforderungen-und-loesungsansaetze> accessed 26 September 2022.

Seaver N, 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems' (2017) 4(2) *Big Data & Society* <https://doi.org/10.1177/2053951717738104> accessed 19 September 2022.

Selbst A and Powles J, 'Meaningful information and the right to explanation' (2017) 7 *International Data Privacy Law* 233.

Sellars A, 'Facebook's threat to the NYU Ad Observatory is an attack on ethical research', *NiemanLab* (29 October 2020) <https://www.niemanlab.org/2020/10/facebooks-threat-to-the-nyu-ad-observatory-is-an-attack-on-ethical-research/> accessed 19 September 2022.

Shane S, 'These are the Ads Russia Bought on Facebook in 2016', *The New York Times* (1 November 2017) <https://www.nytimes.com/2017/11/01/us/politics/russia-2016-election-facebook.html> accessed 19 September 2022.

Shukla S, 'A Better Way to Learn About Ads on Facebook', *Facebook Newsroom* (28 March 2019) <https://newsroom.fb.com/news/2019/03/a-better-way-to-learn-about-ads/> accessed 19 September 2020.

Silva M and others, 'Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook' (2020) *WWW '20: Proceedings of the Web Conference 2020* 224.

Singer N, '"Weaponized Ad Technology": Facebook's Moneymaker Gets a Critical Eye', *The New York Times* (16 August 2018) <https://www.nytimes.com/2018/08/16/technology/facebook-microtargeting-advertising.html> accessed 19 September 2022.

Sobel B, 'A New Common Law of Web Scraping' (2021) 25 *Lewis & Clark Law Review* 147.

Suzor N, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4(3) *Social Media + Society* <https://doi.org/10.1177/205630511878781> accessed 19 September 2022.

Suzor N and others, 'What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation' (2019) 13 *International Journal of Communication* 18.

Swart J, Peters C and Broersma M, 'Shedding light on the dark social: The connective role of news and journalism in social media communities' (2018) 20 *New Media & Society* 4329.

Swisher K, 'Zuckerberg: The Recode interview', *Vox* (8 October 2018) <https://www.vox.com/2018/7/18/17575156/mark-zuckerberg-interview-facebook-recode-kara-swisher> accessed 28 June 2022.

Taekema S, 'Theoretical and normative frameworks for legal research: Putting theory into practice' (2018) *Law and Method* <https://doi.org10.5553/REM/.000031> accessed 19 September 2022.

Tambini D, Leonardi D and Marsden C, 'The privatisation of censorship: self-regulation and freedom of expression', in Damian Tambini, Danilo Leonardi and Chris Marsden, *Codifying cyberspace: communications self-regulation in the age of internet convergence* (Routledge 2008).

Thorburne L, Stray J and Bengagina P, 'What Will "Amplification" Mean in Court?', *Tech Policy Press* (19 May 2022) <https://techpolicy.press/what-will-amplification-mean-in-court/> accessed 19 September 2022.

Thorson K and Wells C, 'Curated flows: A framework for mapping media exposure in the digital age' (2016) 26 *Communication Theory* 309.

Tiku N, 'Facebook has a prescription: More pharmaceutical ads' , *The Washington Post* (4 March 2020) <https://www.washingtonpost.com/technology/2020/03/03/facebook-pharma-ads/> accessed 26 September 2022.

Timmons H and Kozlawska H. 'Facebook's quiet battle to kill the first transparency law for online political ads', *Quartz* (22 March 2018) <https://qz.com/1235363/mark-zuckerberg-and-facebooks-battle-to-kill-the-honest-ads-act/> accessed 19 September 2020.

Tornes A, 'Enabling the future of academic research with the Twitter API', Twitter Developer Blog (21 January 2021) <https://developer.twitter.com/en/blog/product-news/2021/enabling-the-future-of-academic-research-with-the-twitter-api> accessed 26 September 2022.

Trist E and Bamforth K, 'Some social and psychological consequences of the long-wall method of coal-getting' (1954) 4 *Human Relations* 5.

Tufekci Z, *Twitter and Tear Gas: The power and fragility of networked protest* (Yale University Press 2017).

Turton W, 'We posed as 100 senators to run ads on Facebook. Facebook approved all of them', *VICE News* (30 October 2018) <https://news.vice.com/en_ca/article/xw9n3q/we-posed-as-100-senators-to-run-ads-on-facebook-facebook-approved-all-of-them> accessed 19 September 2022.

Turow J, *The Daily You: How the News Advertising Industry is Defining Your Identity and Your Worth* (Yale University Press 2011).

Tutt A, 'An FDA for Algorithms' (2017) 69 *Administrative Law Review* 83.

Twitter, 'Setting the record straight on shadow banning', *Twitter Blog* (26 July 2018) < https://blog.twitter.com/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning> accessed 28 June 2022.

Twitter, Progress Report for the EU Code of Practice on Disinformation (2019) <http://ec.europa.eu/information_society/newsroom/image/document/2019-5/twitter_progress_report_on_code_of_practice_on_disinformation_CF162219-992A-B56C-06126A9E7612E13D_56993.pdf > accessed 26 September 2022.

Twitter, 'How to get certified as a political advertiser', *Twitter Business* (2019) <https://business.twitter.com/en/help/ads-policies/restricted-content-policies/political-content/how-to-get-certified.html> accessed 26 September 2022.

Tworek H and Leerssen P, 'An Analysis of Germany's NetzDG Law' (Transatlantic High Level Working Group on Content Moderation Research Report 2019) <https://www.ivir.nl/publicaties/download/

NetzDG_Tworek_Leerssen_April_2019.pdf> accessed 19 September 2020.

Van Couvering E, 'Is relevance relevant? Market, science, and war: Discourses of search engine quality' (2007) 12 *Journal of Computer-Mediated Communication* 866.

Van Couvering E, *Search Engine Bias: The Structuration of Traffic on the World Wide Web* (PhD thesis, The London School of Economics and Political Science 2010).

Van der Vlist and others, 'API Governance: The Case of Facebook's Evolution' (2020) 8(2) *Social Media+ Society* <https://doi.org/10.1177/20563051221086228> accessed 20 September 2022.

Van Dijck J, Poell T and De Waal M, *The Platform Society: Public Values in a Connective World* (Oxford University Press 2018).

Van Dijck J, 'Users like you? Theorizing agency in user-generated content' (2009) 31 *Media, Culture & Society* 41.

Van Drunen M, Helberger N and Bastian M, 'Know your algorithm: what media organizations need to explain to their users about news personalization' (2019) 9 *International Data Privacy Law* 220.

Van Drunen, Groen-Reijman E, Dobber T, Noroozian A, Leerssen P, Helberger N, De Vreese C and Votta F, 'Transparency and (no) more in the Political Advertising Regulation', *Internet Policy Review* (25 January 2022) <https://policyreview.info/articles/news/transparency-and-no-more-political-advertising-regulation/1616> accessed 6 November 2022.

Van Hoboken J, *Search Engine Freedom: On the Implications of the Right to Freedom of Expression for the Legal Governance of Web Search Engines* (Kluwer International 2012).

Van Hoboken J and Ó Fathaigh R, Regulating Disinformation in Europe: Implications for Speech and Privacy (2019) 6 UC *Irvine Journal of International, Transnational and Comparative law* 9.

Van Til G, 'Zelfregulering door online platforms: een waar wondermiddel tegen online desinformatie?' (2019) 1 *Mediaforum* 13.

Veale M and Ausloos J, 'Researching with Data Rights' [2020] *Technology and Regulation* 136.

Venturini T, 'From fake to junk news: The data politics of online virality', in Didier Bigo, Engin Isin and Evelyn Ruppert (eds), *Data Politics: Worlds, subjects, rights* (Routledge 2019).

Venturini T and Rogers R, '"API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach' (2019) 7 *Digital Journalism* 532.

Vlassenroot E and others, 'Web archives as a data resource for digital scholars' (2019) 1 *International Journal of Digital Humanities* 85.

Wachter S, Mittelstadt B and Floridi L, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 2.

Wachter S, Mittelstadt B and Russell C, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31 *Harvard Journal of Law and Technology* 841.

Wagner B, 'Free Expression? Dominant information intermediaries as arbiters of internet speech', in: Martin Moore and Damian Tambini D (eds.), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (Oxford University Press 2018).

Wagner B and Kuklis L, 'Disinformation, data verification and social media', *Media@LSE* (7 January 2020) <https://blogs.lse.ac.uk/medialse/2020/01/07/disinformation-data-verification-and-social-media/> accessed 20 September 2020.

Wagner B and others, 'Regulating transparency?: Facebook, Twitter and the German Network Enforcement Act' (2020) *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 261.

Waldron J, 'The Rule of 'Law'. *Stanford Encyclopedia of Philosophy* (22 June 2016) <https://plato.stanford.edu/entries/rule-of-law/#ProcAspe> accessed 22 September 2022.

Walter N and others, 'Fact-Checking: A meta-analysis of what works and for whom' (2020) 37 *Political Communication* 350.

Warner M, *Publics and Counterpublics* (Princeton University Press 2002).

Warner R, The Honest Ads Act: primer (US Senate 2017) <https://www.warner.senate.gov/public/index.cfm/the-honest-ads-act > accessed 27 September 2022.

Warren M, 'Accountability and democracy', in Mark Bovens (ed), *The Oxford handbook of public accountability* (Oxford University Press 2014).

Waterson J, 'Obscure pro-Brexit group spends tens of thousands on Facebook ads', *The Guardian* (14 January 2019) <https://www.theguardian.com/politics/2019/jan/14/obscure-pro-brexit-group-britains-future-spends-tens-of-thousands-on-facebook-ads> accessed 20 September 2020.

Waterson J, 'Facebook Brexit ads secretly run by staff of Lynton Crosby firm', *The Guardian* (3 April 2019) <https://www.theguardian.com/politics/2019/apr/03/grassroots-facebook-brexit-ads-secretly-run-by-staff-of-lynton-crosby-firm> accessed 20 September 2022.

Weaver D and others, *The American journalist in the 21st century: US News people at the Dawn of a New millennium* (Routledge 2007).

Webster J, 'Diversity of exposure', in Philip Napoli (ed), *Media Diversity and Localism* (Routledge 2007).

Whittaker J and others, 'Recommender systems and the amplification of extremist content' (2022) 10(2) *Internet Policy Review* <https://doi.org/10.14763/2021.2.1565> accessed 20 September 2022.

Wilman F, *The Responsibility of Online Intermediaries for Illegal User Content in the EU and the US*. (Edward Elgar Publishing 2020).

Winner L, 'Do Artifacts Have Politics?', 109 *Daedalus* 1.

Wisner M, 'Google's Eric Schmidt Responds to Verizon, AT&T Pulling Ads From YouTube', *Fox Business* (23 March 2017) <https://www.foxbusiness.com/features/googles-eric-schmidt-responds-to-verizon-att-pulling-ads-from-youtube> accessed 28 June 2022.

Wodinsky S, 'YouTube's Copyright Strikes Have Become a Tool for Extortion', *The Verge* (11 February 2019) <https://www.theverge.com/2019/2/11/18220032/youtube-copystrike-blackmail-three-strikes-copyright-violation> accessed 11 November 2022.

Wong J, 'One year inside Trump's monumental Facebook campaign', *The Guardian* (28 January 2020) <https://www.theguardian.com/us-news/2020/jan/28/donald-trump-facebook-ad-campaign-2020-election?CMP=Share_iOSApp_Other> accessed 20 September 2020.

Wong J, 'Trump's deluge of Facebook ads have a curious absence: coronavirus', *The Guardian* (26 March 2020) < https://www.theguardian.com/us-news/2020/mar/26/trump-facebook-ads-immigrants-coronavirus> accessed 27 September 2020.

Yeung K, 'Algorithmic regulation: A critical interrogation' (2018) 12 *Regulation & Governance* 505.

YouTube, 'Continuing our work to improve recommendations on YouTube', *YouTube Official Blog* (25 January 2019) <https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html> accessed 26 September 2022.

Zalnierute M, '"Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism' (2021) 8 *Critical Analysis of Law* 39.

Zarsky T, 'Transparent Predictions' (2013) 4 *University of Illinois Law Review* 1503.

Smith M, '"But the data is already public": on the ethics of research on Facebook' (2010) 12 *Ethics in Information Technology* 313.

Zimmer B, '"Techlash": Whipping Up Criticism of the Top Tech Companies', *The Wall Street Journal* (10 January 2019) <https://www.wsj.com/articles/techlash-whipping-up-criticism-of-the-top-tech-companies-11547146279> accessed 24 September 2022.

Zimmer M, '"But the data is already public": on the ethics of research on Facebook' (2010) *12 Ethics in Information Technology* 313.

Zittrain J, 'A History of Online Gatekeeping' (2006) 19 *Harvard Journal of Law and Technology* 253.

Zuckerman E, 'The Case for Digital Public Infrastructure', *Knight First Amendment Institute* (17 January 2020) <https://knightcolumbia.org/content/the-case-for-digital-public-infrastructure> accessed 20 September 2020.

Zuckerberg M, untitled Facebook post, *Facebook*.com (12 January 2018) <https://www.facebook.com/zuck/posts/10104413015393571> accessed 26 September 2022.

Zuckerberg M, 'A Blueprint for Content Governance and Enforcement' (5 May 2021) *Facebook Notes*. <https://www.facebook.com/notes/751449002072082/> accessed 20 September 2020.

Zuckerberg M, 'The Internet needs new rules. Let's start in these four areas', *The Washington Post* (30 March 2019) <https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html> accessed 20 September 2020.

Zuiderveen Borgesius F and others, 'Should we worry about filter bubbles?' (2016) 5(1) *Internet Policy Review* <https://doi.org/10.14763/2016.1.401> accessed 20 September 2020.

Zuiderveen Borgesius F, *Improving Privacy Protection in the Area of Behavioural Targeting* (Kluwer International 2015).

Zuiderveen Borgesius F, 'Behavioural sciences and the regulation of privacy on the internet'. In Anne-Lise Sibony and Alberto Alemanno (eds.), *Nudging and the law: what can EU law learn from behavioural sciences?* (Hart Publishing 2015).

Zuiderveen Borgesius and others, 'Online Political Microtargeting: Promises and Threats for Democracy' (2018) 14 *Utrecht Law Review* 82.

Laws and regulations

European Convention on Human Rights 1950.

Charter of Fundamental Rights of the European Union 2012.

Treaty on the Functioning of the European Union.

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

Regulation 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (P2B regulation).

Regulation 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online.

Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services and amending Directive 2000/31/EC 2020 [COM/2020/825 final] (Digital Services Act).

Proposal for a Regulation of the European Parliament and of the Council on the transparency and targeting of political advertising [COM/2021/731 final] (Political Advertising Regulation).

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (E-Commerce Directive).

Directive 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services in view of changing market realities (AVMS Directive).

Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Copyright Directive).

Staatsvertrag zur Modernisierung der Medienordnung in Deutschland (Federal Republic of Germany)

Loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (Republic of France)

Online Advertising and Social Media (Transparency) Bill 2017 (Republic of Ireland).

Elections Modernization Act 2018-C-76 (Canada).

The Honest Ads Act (proposed), 115th Congress S.1989 (United States).

Case Law

Informationsverein Lentia and others v Austria [1993] ECtHR 13914/88

Jersild v Denmark [1994] ECtHR 15890/89

Verein Gegen Tierfabriken v Switzerland [2001] ECtHR 24699/94

Appleby and Others v United Kingdom [2003] ECtHR 44306/98

Delfi v Estonia [2015] ECtHR 64569/09

Arnett v Kennedy (1974) 416 U.S. 134 (Supreme Court of the United States, 1974)

Campaign Legal Center v. Federal Elections Commission (2020) Case 1:20-cv-00588 (Complaint for declaratory and injunctive relief).

Soft law

Article 29 Working Party, Guidelines on transparency under Regulation 2016/679 (2018) wp260rev.01 < https://ec.europa.eu/newsroom/article29/items/622227/en> accessed 28 June 2022.

Council of Europe, 'Recommendation of the Committee of Ministers to Member States on media pluralism and transparency of media ownership: Guidelines on media pluralism and transparency of media ownership' (2018) CM/Rec(2018)1.

European Commission, Impact Assessment accompanying the Digital Services Act, (2020) SWD(2020) 348 final <https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act> accessed 27 September 2022.

European Data Protection Supervisor, EDPS Opinion 8/2016 on coherent enforcement of fundamental rights in the age of big data (2016) <https://edps.europa.eu/sites/edp/files/publication/16-09-23_bigdata_opinion_en.pdf> accessed 26 September 2022.

European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research (2020) <https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf> accessed 27 September 2022.

EU Code of Practice on Disinformation (European Commission 2018) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> accessed 27 September 2022.

EU Strengthened Code of Practice on Disinformation (European Commission 2022) <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation> accessed 27 September 2022.

Netherlands Ministry of the Interior, 'Response to the Motion for Complete Transparency in the Buyers of Political Advertisements on Facebook' (2019) Kamerstuk 35 078 Nr 26 <https://www.tweedekamer.nl/kamerstukken/detail?id=2019Z03283&did=2019D07045> accessed 27 September 2022.

Parliament of the Netherlands, Motion for Complete Transparency in the Buyers of Political Advertisements on Facebook (2019) Kamerstuk 35 078 Nr 21 <https://www.parlementairemonitor.nl/9353000/1/j9vvij5epmj1ey0/vkvudd248rwa> accessed 27 September 2022.

R

Author contributions

Chapter 3 is co-authored with Jef Ausloos, Brahim Zarouali, Natali Helberger and Claes de Vreese. These co-authors each contributed to the conceptualisation, research and review stages of this project. Brahim Zarouali and Jef Ausloos also contributed to the early drafting stage. The final text is written solely by my hand.

Chapter 4 is co-authored with Tom Dobber, Natali Helberger and Claes de Vreese. Tom Dobber assisted in the research stage and produced the inter-coder reliability scores listed in Table 2. Natali Helberger and Claes de Vreese assisted in the conceptualisation, research and review stages of the project.

For Chapters 1-2 and 5-7 I am the sole author. Of course, these texts also benefited from the helpful guidance of my supervisors Natali Helberger, Claes de Vreese, and Tarlach McGonagle.

A

Summary

This is a dissertation about the transparency of recommender systems in social media governance. Platforms use recommender systems, such as YouTube's Recommended Videos and TikTok's For You, to select and rank content as it is displayed to users. Recommender systems have become important points of control in social media governance, and as their influence grows so have calls for greater transparency and accountability. The goal of this research has been to examine how EU law has gone about this project of' 'opening the black box' of social media recommenders; to interrogate the models of accountability implied in these reforms; and offer alternatives for a more democratically accountable social media governance.

Chapter 1 sets the stage by introducing the main concepts under discussion, formulating my research question, and outlining my research methodology. Social media governance consists in governance *by* and *of* platforms, which governs platformised media-ecosystems and thereby implicates important public interest principles such as media freedom, media pluralism, and the freedom of information and expression. Recommender systems are the algorithmic tools used by these services to order the visibility of content, which they achieve through interrelated processes of curation (selecting items for user relevance and engagement) and moderation (selecting out and sanctioning items which violate applicable rules).

Recommender systems are widely viewed as lacking in transparency, the governance principle that the exercise of power should be knowable to those it affects. Transparency is a precondition for accountability, i.e. the capacity for external actors and institutions – economic, legal, social - to discipline wrongdoing and enforce relevant norms. Transparency regulation therefore prefigures different accountability relationships and power structures by selecting for specific types of information (transparency of what?) and addressing different audiences (transparency for whom?). Transparency regulation is especially challenging in the context of social media recommender systems, due to the scale and complexity of their machine-learning methods (the proverbial algorithmic 'black box') as well as the privacy and security interests at stake in the data being processed.

Accordingly, this dissertation asks: How can EU law regulate the transparency of recommender systems in order to hold online platforms accountable for their role in social media governance? Methodologically, this is a question of normative legal research, which I supplement with insights from the interdisciplinary field of social media studies. Analytically, I take a sociotechnical perspective which highlights the

social context in which technical artifacts are used and given meaning. Normatively, my position, though rooted in fundamental rights law, is based on the principle of 'cooperative responsibility' developed by Natali Helberger, Jo Pierson and Thomas Poell, which seeks to diffuse power and responsibility between a variety of stakeholders, including users and civil society actors, rather than concentrating exclusively on platforms or state regulators as the sole guarantors of the public interest.

Chapter 2 introduces the dominant paradigm of algorithmic transparency regulation for social media recommender systems. It starts by outlining the basic technical and political-economic characteristics of recommender systems, and their increasing importance as an instrument of social media regulation. Particular emphasis is placed on sociotechnical perspectives which highlight the actions of users and platforms as important factors in shaping recommendation outcomes. I then review proposals in European law aiming to enhance recommender system transparency, in terms of the different accountability relationships they pursue: individual disclaimers, regulatory audits and researcher access.

Along these lines, Chapter 2 articulates an initial statement this project's two main positions: First, meaningful transparency about recommender systems cannot focus solely on their algorithms, and must take a broader perspective of recommenders as sociotechnical systems, attentive to the actions of users and platforms in relation to specific content. Second, mechanisms for social accountability should, inasmuch as possible, aim to realise inclusive public resources. Exclusive arrangements with select research partners not only restrict the scalability and potential impact of disclosure, they also risk calling into question the diversity, representativeness and independence of the resulting research – a significant drawback especially in the politically and constitutionally sensitive domain of media policy. Against the EU's technocratic tendency to channel sensitive data toward regulators and trusted experts, I outline possibilities for real-time, outcome-focused and public access to information about platform recommender outcomes and interventions. The subsequent chapters explore these possibilities in greater detail, first as regards content curation and then as regards moderation.

As a case study for public transparency in content curation, Chapters 3 and 4 offer an in-depth analysis of platform ad archives. These tools offer public, machine-readable overviews of advertising distribution via major platform services, along with metadata on their origin and distribution. By focusing on inputs, outcomes and context over algorithmic logics, and by offering public, machine-readable access to non-sensitive data, platform ad archives exemplify the sociotechnical and inclusive approach to transparency put forward in this manuscript.

This case study unfolds over two chapters. **Chapter 3** introduces the phenomenon of ad archives from a governance perspective, describing their legal background and possible accountability functions, as well as the shortcomings in their current self-regulatory implementations. As public tools, I argue, ad archives have the potential to contribute to accountability in several ways. These include legal effects through regulatory monitoring and enforcement, but also social and discursive effects based on the capacity for journalists, academics and other civil society actors to more effectively respond to personalised advertising campaigns. For all these potential benefits, I also discuss the limitations and shortcomings in ad archives' present implementation. These problems include the selection of political ads included in the archive; the dubious quality of platform data about user identity and analytics; and the crucial omission of any information about ad targeting mechanisms. On this basis I provide several proposals for public regulation to improve ad archives. (After the time of writing, the Digital Services Act would take up this endeavour in Article 39).

Chapter 4 complements my theoretical account of ad archive accountability with an original empirical investigation into the usage of an ad archive by journalists (specifically: the tool launched by Facebook, the 'Ad Library'). These journalistic practices are mapped through content analysis of news reporting which references the ad archive, combined with interviews with relevant journalists. This research confirms that journalists in various countries have made repeated use of the ad archive for reporting purposes. Such usage was relatively more common in the US and UK, compared to the Netherlands and Germany where political advertising is less prevalent. The most specialised journalists still relied on independent scraping to enrich and verify ad archive data, underscoring the continued importance of these independent collection methods. In terms of accountability, our evidence suggests ad archives have catalysed 'hard' legal accountability through investigative watchdog reporting drawing attention to wrongdoings, including potential violations of legal and contractual norms, alongside 'soft' discursive accountabilities associated with more generic campaign reporting about microtargeted messaging strategies.

Chapter 5 shifts focus from content curation to content moderation; how platforms intervene in ranking outcomes and suppress specific items in order to enforce applicable rules. This relatively novel technique of visibility reduction, I argue, is less transparent than conventional methods such as content takedown or account suspension, since its outcomes are obscured by the personalised volatility of recommender systems. Unless these measures are expressly notified to users, they remain invisible and result in what has come to be known in popular and academic discourses as 'shadow banning' – sanctions that are unnoticeable to the affected user.

I discuss the justifications that platforms offer for shadow banning, questioning their narrative of security and efficacy and highlighting their unstated political interests in secrecy. I then discuss how content moderation due process rules, laid down in Article 14 and 17 of the new Digital Services Act (DSA), can be read as a prohibition on shadow banning, with limited exceptions. In implementing these notice rights, I argue, an important technical challenge will be to define visibility reductions as a category of moderation sanctions distinct from the routine operations of recommender curation.

Chapter 6 draws together insights from the foregoing chapters to propose a reframing of data access regulation for social media recommender systems, away from algorithmic transparency and toward platform observability. In this chapter I show how the principle of observability, first proposed by Bernhard Rieder and Jeannette Hoffman, aligns with the inclusive and sociotechnical forms of data access regulation advocated throughout this manuscript. I then use the principle of observability to review the regulatory reforms proposed by the Digital Services Act (DSA), and particularly its data access framework laid down in Article 40, which can be understood an early attempt to surpass the algorithmic explanation paradigm and to start regulating for and with observability. In doing so, however, the DSA surfaces important challenges. Regulating *for* observability faces trade-offs between inclusiveness and depth of access and line-drawing problems around the publicness of user content. And in regulating *with* observability, tensions arise between observability's direct role in law enforcement and its more indirect roles in knowledge production and public discourse. I argue for a loose coupling between observability and regulatory enforcement, and against the tendency to reduce data access to mere compliance monitoring.

Chapter 7 offers general conclusions and closes with an outlook on the future of observability regulation for social media. I discuss how the case studies of ad archives and shadow banning safeguards both shed light on different aspects of recommender transparency, respectively for content curation and moderation. What makes these reforms distinctively meaningful compared to generic algorithmic explanation duties is that they focus on context-specific decisions and outcomes – the prior questions of *what* content being curated and moderated, as a precursor to any meaningful discussion as to *why*. Both of these methods can still be extended in future. Ad archives raise the question whether similar resources can also be developed for other categories of (organic) public content. And the DSA's shadow banning safeguards for uploaders still leave unaddressed whether visibility reductions should also be made known to broader publics.

At its most general level, therefore, this dissertation's recommendation is for law and policymakers to take up the task of regulating inclusive observability of recommender outcomes in social media governance. Compared to algorithmic explanation, this disclosure model is less complex and less sensitive, and therefore permits more public and inclusive access toward non-institutional actors such as politicians, activists and journalists. In this way observability acts as an essential means of social accountability in social media governance, and as a catalyst for legal ordering; by enabling publics, across personally curated flows, to see what others are seeing.

Samenvatting

Dit is een proefschrift over de transparantie van aanbevelingssystemen in sociale media. Platforms gebruiken aanbevelingssystemen, zoals YouTubes Recommended Videos en TikToks For You, om de inhoud die aan gebruikers wordt getoond te selecteren en rangschikken. Aanbevelingssystemen zijn verworden tot belangrijke instrumenten van regulering in sociale media governance, en naar mate hun invloed toeneemt stijgt ook de vraag naar transparantie en toerekenbaarheid ('accountability') in deze systemen.[1] Het doel van dit proefschrift is om te onderzoeken hoe het EU-recht dit project van transparantieregulering aanpakt; om de toerekenbaarheidsrelaties en machtsstructuren die met deze transparantieregels gepaard gaan te verhelderen en kritisch te bevragen; en alternatieven aan te rijken voor een democratisch toerekenbare governance van en door sociale media.

Hoofdstuk 1 begint met een inleiding tot de belangrijkste concepten van dit proefschrift, gevolgd door de onderzoeksvraag en methodologie. Sociale media governance beschrijft een governance *van* en *door* platforms, die een geplatformiseerd media-ecosysteem reguleren en daarmee publieke belangen treffen zoals mediavrijheid, mediapluriformiteit, en de vrijheid van informatie en meningsuiting. Aanbevelingsystemen beheren op deze diensten de zichtbaarheid van gebruikersinhoud, door een combinatie van inhoudscuratie (het selecteren van inhoud op relevantie en engagement) en inhoudsmoderatie (het identificeren en sanctioneren van inhoud die de regels van het platform schendt).

Aanbevelingssystemen worden bekritiseerd wegens een gebrek aan transparantie, oftewel het beginsel dat de uitoefening van macht kenbaar moet zijn aan diegenen die het treft. Transparantie is een voorwaarde voor toerekenbaarheid, ofwel het vermogen van externe actoren en instellingen – economisch, juridisch, of sociaal – om wangedrag te tuchtigen en relevante normen te handhaven. Transparantieregulering loopt daarom vooruit op bepaalde toerekenbaarheidsrelaties en machtsstructuren, door bepaalde soorten informatie wel of niet af te dwingen (transparantie waarvan?) en door bepaalde actoren wel of niet te addresseren (transparantie voor wie?). Transparantieregulering is bijzonder uitdagend in de context van aanbevelingssystemen, vanwege de schaal en complexiteit van hun machine-learning algoritmes – de zogehete 'black box' – alsook vanwege de privacy- en securitybelangen die gepaard gaan met deze data.

1 De term 'governance', zoals ik het gebruik, kent geen geschikte vertaling in het Nederlands. Het is vergelijkbaar met regulering, maar verwijst uitdrukkelijk naar regulering door private partijen zoals platforms en naar niet-juridische reguleringstechnieken zoals technologische regulering.

Dit proefschrift stelt daarom de vraag: hoe kan de EU de transparantie van aanbevelingssystemen reguleren om online platforms toerekenbaar te maken voor hun rol in sociale media governance? Methologisch is dit een normatief juridisch onderzoek, waarbij ook inzichten uit het interdisciplinaire onderzoeksveld sociale media studies worden betrokken. Analytisch ga ik uit van een sociotechnisch perspectief dat de nadruk legt op de sociale context waarin technische artefacten toepassing vinden en betekenis krijgen. Normatief berust mijn kritiek zich op de bescherming van fundamentele rechten, bezien door de lens van 'cooperative responsibility' ontwikkeld door Natali Helberger, Jo Pierson en Thomas Poell. Dit ideaal beoogt om macht en verantwoordelijkheid in platform governance uit te spreiden over verscheidene stakeholders, waaronder gebruikers en het maatschappelijke middenveld, in plaats van deze uitsluitend in platforms of toezichthouders te concentreren als de enige behoeders van publieke belangen.

Hoofdstuk 2 introduceert het dominante paradigma van algoritmische transparantie bij de regulering van aanbevelingssystemen op sociale media. Eerst beschrijft het de technische en politiek-economische kenmerken van aanbevelingssystemen, en hun toenemende rol in de regulering van sociale media. In het bijzonder benadruk ik hierbij sociotechnische literatuur die beschrijft hoe de handelingen van gebruikers en platformbedrijven een vergaande invloed uitoefenen op aanbevelingsuitkomsten. Dan beschrijf ik hoe het Europese recht in vercheidende beleidsvoorstellen beoogt om de transparantie van deze systemen te reguleren, aan de hand van de verschillende toerekenbaarheidsrelaties die zij nastreven: toelichtingen en bijsluiters voor eindgebruikers, audits en rapportageverplichten voor toezichthouders, en toegangspartnerschappen voor onderzoekers.

In dit licht worden de de twee belangrijkste uitgangspunten van dit onderzoeksproject gearticuleerd: een *sociotechnisch* en *inclusieve* benadering tot transparantie. Het sociotechnische perspectief, ten eerste, behelst dat transparantiebeleid zich niet alleen om aanbevelingsalgoritmes dient te bekommeren, maar juist een breder perspectief dient te ontwikkelen op de aanbevelingspraktijk als sociotechnisch systeem, met aandacht voor het handelen van de gebruiker en het platform in relatie tot specifieke inhoud. Een inclusieve aanpak bepleit dat transparantiemaatregelen, voor zover mogelijk gelet op privacybelangen, idealiter publiek toegankelijk zijn. Exclusieve arrangementen waarbij specifieke actoren worden geselecteerd voor toegang, beperken niet alleen de algemene schaalbaarheid en mogelijke impact van transparantieoplossingen, maar kunnen ook de diversiteit, representativiteit en daarmee legitimiteit van het daaruit voortvloeiende onderzoek in twijfel doen trekken – zeker in het politiek gevoelige domein van mediaregulering een wezenlijke

beperking. Tegen de technocratische tendensen van Europees beleidsmakers om gevoelige data exclusief aan toezichthouders en vertrouwde experts voor te behouden, bespreek ik mogelijkheden voor real-time, uitkomstgerichte en publieke toegang tot data over aanbevelingssystemen.

Als case study voor deze soort publieke transparantie, bieden Hoofdstuk 3 en 4 een gedetailleerde bespreking van platformadvertentie-archieven. Deze faciliteiten bieden publieke, digitale overzichten van advertenties op een bepaald platform, gepaard met relevante metadata over hun origine en distributie. Deze recente beleidsontwikkeling typeert de sociotechnische en inclusieve transparantiemethode die centraal staat in dit proefschrift, ten eerste door zich te richten op de inputs, outputs en context van gepersonaliseerde distributie, en niet slechts de algoritmische logica; en ten tweede door deze informatie publiek toegankelijk te maken.

Deze case study is verdeeld over twee hoofdstukken. **Hoofdstuk 3** beschrijft advertentie-archieven als een nieuw fenomeen in platform governance. Advertentie-archieven kunnen op verschillende manieren gebruikt worden en tot toerekenbaarheid bijdragen. Te denken valt aan juridische effecten door het vergemakkelijken van toezicht en handhaving, maar ook sociale en discursieve effecten bestaande in de capaciteit voor journalisten, academici en andere maatschappelijke actoren om te kunnen reageren op gepersonaliseerde reclamecampagnes. Daartegenover staat wel kritiek uit de onderzoekspraktijk over de beperkingen en tekortkomingen in het huidige ontwerp van advertentie-archieven. Belangrijke punten van kritiek zijn de onduidelijke selectiemethoden voor het wel of niet archiveren van bepaalde advertenties; de dubieuze kwaliteit van de informatie over gebruikersgedrag en -identiteit; en het cruciale gebrek aan enige informatie over targetingcriteria. In dit licht bespreek ik enkele voorstellen om advertentie-archieven te verbeteren middels publiek toezicht.

Hoofdstuk 4 ondersteunt het theoretische argument voor advertentie-archieven met origineel empirisch onderzoek naar het gebruik door journalisten van één zulk archief (te weten, het archief van Facebook, hun zogehete 'Ad Library' of 'Advertentiebibliotheek'). Deze journalistieke praktijk wordt in kaart gebracht door een inhoudsanalyse van nieuwspublicaties die naar dit archief verwijzen, gecombineerd met interviews met betrokken journalisten. Dit onderzoek bevestigt dat journalisten in verschilllende landen meermaals gebruik hebben gemaakt van de ad archive in hun verslaggeving. Zulk gebruik kwam vaker voor in de VS en het VK, vergeleken met Nederland en Duitsland (waar politieke advertenties dan ook minder voorkomen). Tegelijkertijd hechten gespecialiseerde onderzoeksjournalisten op dit gebied nog

wel groot belang aan onafhankelijke datavergaringsmethoden ('data scraping'), om archiefdata aan te vullen en te verifiëren. Qua toerekenbaarheid geeft dit onderzoek aan dat advertentie-archieven 'harde' juridische toerekenbaarheid hebben gecatalyseerd met onderzoeksjournalistiek die de aandacht vestigt op wanpraktijken, waaronder schendingen van wettelijke en contractuele normen, maar daarnaast ook 'zachte' discursieve toerekenbaarheid bestaande in algemene politieke verslaggeving over gepersonaliseerde advertentiecampagnes.

Hoofdstuk 5 verlegt de focus van inhoudscuratie naar inhoudsmoderatie, en de vraag hoe platforms interveniëren in aanbevelingssystemen om bepaalde inhoud te onderdrukken en daarmee hun inhoudsbeleid te handhaven. Deze relatief nieuwe techniek van zichtbaarheidsvermindering ('visibility reduction') is minder transparant dan conventionele moderatietechnieken zoals inhoudsverwijdering of accountblokkade; zulke sancties zijn moelijk waarneembaar temidden van het gepersonaliseerde en volatiele gedrag van aanbevelingssystemen waarop zij ingrijpen. Tenzij expliciet vermeld aan de gebruiker, blijft de sanctie dus geheim – ook wel bekend als een 'shadow ban'. Platforms rechtvaardigen shadow bans op grond van veiligheid en effectiviteit, maar ik trek dit narratief in twijfel door te wijzen op de politieke belangen die ook bij dit vraagstuk spelen; hoe geheimhouding ook kan dienen als strategie om juridische en sociale toerekenbaarheid bij inhoudsmoderatie te ontvluchten. Vervolgens bespreek ik nieuwe wetgeving op dit gebied, de Verordening Digitale Diensten, en hoe diens procedurele waarborgen voor inhoudsmoderatie strekken tot een verbod op shadow banning, met beperkte uitzonderingen. Een belangrijke uitdaging voor het handhaven van deze regels blijft de technische definitie van zichtbaarheidsvermindering, als categorie van sancties, en hoe deze kunnen worden onderscheiden van meer routineuse aspecten van inhoudscuratie.

Bij wijze van synthese stelt **Hoofdstuk 6** voor om een nieuw frame te adopteren voor de regulering van sociale media aanbevelingssystemen, niet langer als een kwestie van algoritmische transparantie maar voortaan als een kwestie van platformobserveerbaarheid ('platform observability'), afkomstig uit het werk van Bernhard Rieder and Jeannette Hoffman. In dit hoofdstuk bespreek ik hoe het principe van observeerbaarheid aansluit bij het inclusieve en sociotechnische perspectief van dit proefschrift. Waar de transparantiemetafoor bijvoorbeeld uitgaat van een blik naar *binnen* (en dus het inspecteren van algoritmische 'black boxes'), biedt observeerbaarheid een breder blikveld (niet alleen de black box naar binnen maar ook met oog voor externe relaties en context). In termen van observeerbaarheid bespreek ik vervolgens de nieuwe Verordening Digitale Diensten. In het bijzonder diens regels voor datatoegang door onderzoekers in Artikel 40 kunnen gezien worden als een

eerste stap in deze overgang van algoritmische transparantie naar observeerbaarheid. Daarmee worden ook uitdagingen zichtbaar. Reguleren *voor* observeerbaarheid vereist een afweging van inclusiviteit tegen diepgang, en het trekken van grenzen tussen publieke en niet-publieke communicatie op sociale media. En bij het reguleren *met* observeerbaarheid onstaat er een spanningsveld tussen een directe rol in handhaving en een indirecte rol in kennisproductie en publiek debat. Ik pleit voor een losse koppeling tussen observeerbaarheidsregulering en handhaving, en tegen de tendens om de rol van platformdata te reduceren tot toezicht op naleving van de wet.

Hoofdstuk 7 sluit af met algemene conclusies. De case studies van advertentie-archieven en shadowbanningregels bieden allebei inzage in verschillende aspecten van het aanbevelingsproces, respectievelijk als inhoudscuratie en als inhoudsmoderatie. Wat deze vormen van informatie betekenisvol maakt, vergeleken met conventionele plichten tot algoritmische transparantie, is dat zij zich richten op uitkomsten en beslissingen in relatie tot specifieke inhoud: de essentiële voorvraag *welke* inhoud wordt gecureerd en gemodereerd, voorafgaand aan het *waarom*. Beide methodes zouden in de toekomst nog verder kunnen worden uitgebreid: advertentie-archieven werpen de vraag op hoe dergelijke principes kunnen worden uitgebreid naar andere ('organische') publieke inhoud op sociale media. Bij shadow banning daarentegen resteert juist de vraag wanneer deze maatregelen niet alleen aan de betrokkene maar ook naar andere partijen kenbaar dienen te worden gemaakt.

De belangrijkste aanbeveling van dit proefschrift is daarom dat beleidsmakers deze kansen aangrijpen om sociale media aanbevelingssystemen op inclusieve wijze observeerbaar te maken. Vergeleken met algorithmische uitleg is observeerbaarheid minder complex en minder gevoelig, en daarom vatbaar voor brede publieke toegang door politici, activisten, journalisten en andere non-institutionele actoren. Zo opereert observeerbaarheid als essentieel middel van sociale toerekenbaarheid in sociale media governance, en als catalysator van bindend juridisch toezicht; door de gepersonaliseerde informatiestromen van sociale media te doorklieven, en het publiek in staat te stellen om te zien wat anderen zien.

Appendix I: Content Analysis Protocol

Journalistic use of the Facebook Ad Library

This protocol addresses journalistic use of Facebook's Ad Library for print media publications. The sampled articles were selected based on keyword searches in the LexisNexis database:

Keywords: "Facebook" AND "ad archive" OR "ad library"

Period: 1 May 2018 – 10 July 2020

Publication type: Newspapers + Magazines & Journals

Location: United States + United Kingdom + Germany + The Netherlands

NON-METACOVERAGE

Does the article refer to ad information sourced from the Ad Library, for purposes that are not merely illustrative of the Ad Archive's affordances?

Filter out reporting that merely describes the Ad Archive *as a new phenomenon,* rather than utilizing the Ad Archive *as a resource for other newsworthy information.*

N.B: References to the ad archive can be made in plaintext (i.e. 'Facebook's Ad Library shows that [x]'; 'According to Facebook's Ad Library, [x]' , but also on the basis of recognizable screenshots (see examples below) or URL links (the domain 'http://facebook.com/ads/library').

1. The article does not use Ad Library data, or this usage is merely illustrative.

2. The article does not use Ad Library data, or this usage is merely illustrative. The article uses Ad Library data, and this usage is not merely illustrative.

If 1, continue coding the following questions. If 0, proceed to next article without coding the following questions.

POLITICAL ADS

Determine whether the advertising cited from Ad Library is commercial or political. For purposes of this analysis, commercial ads are ads which aim to encourage the sale of goods or services. Political ads are non-commercial ads. In addition, if the article states that Facebook has designated the ad 'political', it should be coded as such.

For example, ads related to President Trump's hotels in Scotland would *prima facie* qualify as commercial, since they relate to the sale of goods or services. However, if the article states that Facebook has classified these ads as 'political', then they should be coded as such.

1. The article refers to political ads in the Ad Library

2. The article refers to commercial ads in the Ad Library

3. The article refers to both political and commercial ads in the Ad Library

4. The commercial/political status of the ads referred to is unspecified or unclear

WRONGDOING

Does the article describe the ads cited from the Ad Library as potentially harmful, illegal, unethical, or otherwise involving potential wrongdoing? The allegation may originate from a third party source or quote, but it must in any case be explicit.

N.b.: Wrongdoings must relate to the ads cited from the ad library (coded under the previous question). Without an explicit connection to the material cited from the ad archive, generic references to wrongdoing in other ads or practices should not be taken into consideration.

0 No.

1 Yes.

WRONGDOING CATEGORY

If 'yes' to the above question, which of the following categories of potential harms does the alleged wrongdoing relate to? Specify whether this alleged wrongdoing is referenced in the article's headline.

Content of the advertisement: e.g. falsehoods or half-truths; misleading content; offensive content.

1. No.
2. Yes, but not in the article headline.
3. Yes, also in the article headline.

Personalization practices: e.g. discriminatory targeting of ads; risk of excluding or marginalizing certain audiences; chance of manipulating or deceiving audiences through targeting choice.

1. No.
2. Yes, but not in the article headline.
3. Yes, also in the article headline.

Identity of the ad buyer & origin of funds: e.g. related to legitimacy of participation by ad buyers – e.g. due to involvement of foreign entities; campaign finance considerations; astroturfing; the misleading, deceptive or opaque identity of ad buyers & the origin of their funds.

1. No.
2. Yes, but not in the article headline.
3. Yes, also in the article headline.

WRONGDOING NORM?

Does the article claim that the wrongdoing described above may potentially violate / have violated applicable laws?

1. No.
2. Yes.

Does the article claim that the wrongdoing described above may potentially violate / have violated Facebook's Terms of Service?

1. No.
2. Yes.

Appendix II: Supplemental keyword testing

Our content analysis is based on LexisNexis searches with the keywords <"Facebook" AND "Ad Library" OR "Ad Archive">. In theory, our keywords have the potential to overstate the prevalence of metacoverage. After all, one might predict that metacoverage describing the Ad Library is also more likely to explicitly reference this tool by name, compared to substantive usage where its name may be of secondary importance to its contents. Since the relative prevalence of metacoverage is an important finding in our paper, we have run additional tests to check for such a bias.

Using alternative keywords, we were able to demonstrate that our keywords did not significantly bias our findings as regards metacoverage. In short, we found that metacoverage is at least as likely as substantive usage to use non-standard referencing that escapes our initial keywords. Below we provide a more detailed explanation of these tests:

In LexisNexis, we ran additional searches for the United States keeping variables identical (time period, publication type) except for keywords. We selected the United States since it is the largest region in our sample and because its share of metacoverage (62%) in our original analysis is closest to the overall average of 60%. We searched for the following alternative keywords:

Sample I: <"Facebook" AND "political ads"> , and

Sample II: <"data provided by Facebook" OR "data made available by Facebook" OR "data published by Facebook">.

Sample I returned 996 results, Sample II only 3.

As regards sample II we can be brief, since it only returned three results. Evidently these phrasings are not as common as one might perhaps expect, and their omission has not significantly affected our content analysis. The three results we found included one duplicate, and the two remaining articles did not concern political advertising, much less the Ad Library.

Sample I is more informative and more complex. We discuss our analysis in greater detail below.

Given the large number of results (996), we selected a random sample of 50 articles with the help of Google's Random Number Generator tool. We read each article in this sample and tried to ascertain whether it used data from the ad library, and classified them as follows (see also the attached code sheet):

In this sample we found 7 cases of **metacoverage**. One of these articles contained a verbatim reference to the 'ad archive', and had already been coded as metacoverage in our previous analysis. Six other articles referred to the Ad Library expressly but with slightly different terminology, such as "public database" or "searchable archive"—a possibility we also discuss in our paper. However, in each instance, these articles all constituted metacoverage as they did not cite any data from the tool.

By contrast, we found only 3 articles where **ad archive usage** is apparent. The first case attributed ad data to Facebook in general, rather than the Ad Library in particular: "According to Facebook data, the Trump campaign spent $21.25 million", etc. Given the content and context, we consider it likely that this data originates from the Ad Library. The second case, a USA Today's investigation into false positive detection of political ads, does not specify its methodology but does contain screenshots that are evidently taken from the Ad Library. Finally, the third case refers to an investigation by 60 Minutes that relies on the Ad Library; the ad archive is only referenced in the underlying source and not in the sampled article that references it, and so it did not occur in our original search.

In addition, we encountered **three inconclusive cases** where we cannot be certain, as they contain claims about Facebook advertising without any clear source. Overall, we consider Ad Library usage unlikely in these cases, though it cannot be ruled out entirely, for reasons we explain below. These articles all describe controversies around specific ad campaigns, including those of Elizabeth Warren, California gubernatorial candidate Adriel Hampton, and Falun Gong news outlet the Epoch Times. These articles rely primarily on interviews and public statements from Facebook and the campaigns involved. At certain points, these articles describe the ad content and/or spending, without discrete sources or attributions for these claims, in ways that may conceivably draw on the Ad Library. For instance, the Epoch Times article cites Facebook spend figures without a discrete source, but later on cites their YouTube ad spend credited to a "a person familiar with its spending". The Facebook figure could conceivably be drawn from the Ad Library, but in our judgement the more plausible interpretation is that it comes from the same anonymous source as the YouTube figure. Comparable reasoning applies for the other two articles. In addition it is worth noting that these articles do not use screenshots of the ads at issue or cite other relevant

Ad Library data such as demographics, as one might expect from a journalist who had been able to track down the ads in such a way -- and as we commonly see in the substantive usage from our original content analysis. In such cases, therefore, we can only conclude that Ad Library usage is conceivable, but not particularly probable.

In any event, the total apparent usage (3) + inconclusive cases (3) is still fewer than the meta-coverage in our new sample (7). So even if we treat these inconclusive cases as conservatively as possible and assume the Ad Library was in fact used in all instances, metacoverage is still more prevalent. In our view a more realistic estimate, for the reasons cited above, would be a metacoverage-to-usage ratio of 7:3, or 70% metacoverage.

Recalling that our original content analysis yielded a metacoverage percentage of 58% overall and 62% for the US in particular, these results are therefore in line with our findings. Depending on how uncertainties are treated, the new keywords could lead to either a somewhat lower ratio, or, more plausibly, an even higher ratio. Certainly a clear bias towards metacoverage is not evident, and even the most conservative interpretation of this test still supports our finding that meta-coverage outnumbers substantive usage.