REGULATING DISINFORMATION ON SOCIAL MEDIA PLATFORMS

A Defence of the Meta-regulatory Framework

Disinformation on social media platforms has become a serious problem in recent years. This article argues that due to addictive design and the gatekeeping power of social media companies, self-regulation is ineffective because it usually devolves into either over-regulation or reluctant regulation. The former is the practice of overblocking, removing all suspicious content with no concern for user rights, while the latter denotes the unwillingness to regulate false content in order to profit from higher user involvement. This article argues that governments can enforce the regulation of disinformation effectively without falling into the trap of over-regulation under a meta-regulatory framework. Meta-regulation has two policy objectives: one is to enforce platforms to regulate disinformation effectively, and the other is to prevent platforms from over-regulating user speech. When platforms regulate disinformation, they need to meet certain substantive and procedural requirements. These obligations can create incentives for platforms to establish effective processes to identify and remove "disinformative" content, as well as provide users procedural protection. Given that disinformation is amplified by the gatekeeping power of platforms, which has created an imbalance between users and platforms, public authority is justified to adopt certain measures, such as a meta-regulatory framework, to protect the fundamental right of users. This article then responds to potential concerns about this framework, such as the increasingly asymmetrical power of platforms over users and the risk of delegating public authority.

YANG Shao-Kai
LLB, LLM (National Taiwan University);
Attorney-at-Law (Taiwan);
Attorney, Oz & Goodwin Global Law Firm.

I. Introduction

1 Social media platforms have played an influential role in communication systems in recent years. These platforms provide users

with a place for online communication and information sharing.[1] During the COVID-19 pandemic, social media platforms were flooded with disinformation, which is defined in this article as false information "deliberately created to harm or mislead a person, social group, organisation, or country".[2] Disinformation about COVID-19 has rapidly proliferated, causing negative impact on public health policy.[3] Due to the far-reaching consequences of disinformation, finding best regulatory practices has become a central issue to governments and social media platforms. So far, these platforms have implemented several self-regulatory methods to counter disinformation, such as working with third-party organisations to review actions. However, this article will argue that due to the addictive design of platforms and the gatekeeping power platforms exercise, most current regulatory methods are flawed. A better model will be proposed later in this article.[4]

2 In Part II.A, this article first illustrates the addictive design of social media platforms and then discusses the need for private governance. In Part II.B, this article argues that from a communication systems perspective, the private governance of platforms provides the gatekeeping power to influence the production, expression and dissemination of speech. Nevertheless, the exercise of such gatekeeping power also constitutes part of the disinformation problem. That is to say, one cannot tackle disinformation without taking platforms into consideration.

3 In Part III.A, this article discusses existing self-regulatory efforts to tackle disinformation on various platforms, which are not sufficiently effective. From the perspective of addictive design, platform self-regulation tends to devolve into reluctant regulation; from the

---

1 Sofia Grafanaki defined social media platforms as private entities with two distinct roles in the systems of information flow. First, they host online public expression and second, they provide navigation and delivery of the digital content of others: see Sofia Grafanaki, "Platforms, the First Amendment and Online Speech Regulating the Filters" (2018) 39 Pace L Rev 111 at 116.

2 In contrast, misinformation is information that is false, but not created with the intention of causing harm. See Claire Wardle & Hossein Derakhshan, *Information Disorder Toward an Interdisciplinary Framework for Research and Policymaking* (Council of Europe, 27 September 2017) at pp 20–22 <https://rm.coe.int/information-disorder-report-2017/1680766412> (accessed 10 December 2019).

3 Ingrid Volkmer, *Social Media and Covid-19: A Global Study of Digital Crisis Interaction Among Gen Z and Millennials* (Wunderman Thompson, 1 December 2021) at p 5 <https://www.who.int/news-room/feature-stories/detail/social-media-covid-19-a-global-study-of-digital-crisis-interaction-among-gen-z-and-millennials> (accessed 8 December 2021).

4 For the purposes of this article, it will be assumed that it is possible for the government to regulate disinformation without violating freedom of speech. This article will not focus on whether regulation of disinformation can be justified, but on how disinformation can be regulated effectively.

perspective of freedom of speech, while platforms might have their own ways of regulating disinformation, they might not protect users' freedom of speech when they overly regulate content tagged as disinformation.

4 In Part III.B, this article examines the drawbacks of self-regulation and proposes meta-regulation as a solution. It further argues that meta-regulation is justified in terms of the gatekeeping power of platforms.

5 In Part IV.A, this article applies meta-regulation to tackle disinformation problems. Within a meta-regulatory framework, governments would identify disinformation and develop plans with social media platforms to enforce regulations without overly restricting users' right of speech. In this way, platforms can retain their business models as long as they meet public objectives.

6 In Part IV.B, this article discusses some concerns and risks of meta-regulation and provides some solutions.

II. Social media and disinformation

A. Addictive design and the private governance of platforms

7 In Part I, this article briefly discussed addictive design and its profit-seeking business model based on attracting user attention. Platforms enforce rules to guarantee profitability. The enforcement of these rules is called private governance. This Part highlights the need for private governance.

(1) Addictive design and private governance

8 Whenever users spend time on platforms, they are creating data when they post new content, or when they like or click on a link. While services on platforms seem to be free, platforms are profiting from this data: they collect, store and sell data to advertisers or business partners.[5] As Tim Wu pointed out, social media platforms have become attention-brokers in the 21st century.[6] As attention-brokers, platforms try to maximise the amount of time and attention people spend on them.[7] The more time and attention users spend, the more data they generate and the

5 Jack M Balkin, "The First Amendment in the Second Gilded age" (2018) 66 Buff L Rev 979 at 990–991.

6 Tim Wu, "Blind Spot: The Attention Economy and The Law" (2019) 82 Antitrust LJ 771 at 783.

7 Denis McQuail & Mark Deuze, *McQuail's Media & Mass Communication Theory* (SAGE Publications, 7th Ed, 2020) at p 96; Jack M Balkin, "Free Speech in the

(cont'd on the next page)

more profits platforms can make from advertising.[8] Therefore, one can say that the business model is driven by the data users create.

9 Denis McQuail and Mark Deuze suggest that there are three features of this business model: first, communication exists only in present, the past does not matter, and the future matters only when it is an amplification of the present; second, attention is scarce, and attention-gaining is a zero-sum game; third, attention-gaining is an end in itself and is value-neutral in the short term.[9] While one can be positive that user attention is limited and platforms seek to increase the time users spend on them, platform attention-gaining efforts are in line with corporate interest and are thus not value-neutral but profit-seeking.

10 As a result of operating on addictive design, platforms need to establish mechanisms for governance to regulate posted content to reach their economic goals. That becomes the private governance of platforms.

11 The concept of private governance used in this article is borrowed from Kate Klonick. She defined private governance as a set of new governance models that identify the interplay between users and platforms, such as dynamic and iterative law-making processes, norm-generating individuals, and the convergence of processes and outcomes.[10] Klonick views these platforms as new governors, for they developed elaborate bureaucracies and had taken their role as community governors to judge when and whether to remove posts or suspend users.[11] They have centralised bodies, established sets of rules, and *ex ante* and *ex post* procedures for adjudication. The ways in which they decide how to self-regulate themselves reflects the norms of a community.[12]

12 The basis of private governance is the terms of service between the users and platforms.[13] Most of these terms, as Koltay pointed out, are

---

Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1192.

8 Tim Wu, "Blind Spot: The Attention Economy and The Law" (2019) 82 Antitrust LJ 771 at 791; Tim Wu, "Is the First Amendment Obsolete?" (2018) 117 Mich L Rev 547 at 555.

9 Denis McQuail & Mark Deuze, *McQuail's Media & Mass Communication Theory* (SAGE Publications, 7th Ed, 2020) at pp 96–97.

10 Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harv L Rev 1598 at 1616–1617.

11 Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harv L Rev 1598 at 1616–1617.

12 Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harv L Rev 1598 at 1616-1617.

13 Marjorie Heins, "The Brave New World of Social Media Censorship" (2013–2014) 127 Harv L Rev 325 at 325–326.

unilaterally formulated by platforms and can be added or amended at any time. Platforms also often use vague terms to reserve the right of final decision.[14] The terms of service affect user speech on the platform, and the contractual power of the terms also shield platforms from requirements of freedom of speech because users have agreed to their terms.

13 In the following section, this article will argue why private governance is inevitable for platforms and why this discussion is needed if we want to clarify the relationship between disinformation and platforms.

(2) The need for private governance of speech

14 The private governance of speech is inevitable because platforms are under pressure to moderate content. These pressures include the need to maintain a comfortable space for users and the need to comply with established regulations or law (avoid cybercrimes, complying with intellectual property rights, *etc*). These factors are important in the view of economic considerations: in order to make profits, platforms need not only to attract more users to join membership but also to make users feel safe, so that users will spend more time and generate more data via their platforms.[15] In other words, social media platforms need to implement private governance systems to curate and build an attractive environment to engage users if they want to make profit.[16]

15 Due to the sheer number of users, private governance is largely implemented by algorithms.[17] Facebook (now Meta), for example, provides users personalised content with algorithms to retain user attention. The curation process determines what users can say on the platform, what posts can be seen, and even who can stay on the platform.[18] As Grafanaki stated, the private governance enforced by platforms includes content moderation policies that determine whether the content can be hosted

14 András Koltay, *New Media and Freedom of Expression Rethinking the Constitutional Foundations of the Public Sphere* (Bloomsbury Publishing, 2019) at p 183.

15 Benjamin F Jackson, "Censorship and Freedom of Expression in the Age of Facebook" (2014) 44 NM L Rev 121 at 127–131; Jack M Balkin, "Free Speech Is a Triangle" (2018) 118 Colum L Rev 2011 at 2022–2023.

16 Jack M Balkin, "The First Amendment in the Second Gilded age" (2018) 66 Buff L Rev 979 at 997; Jack M Balkin, "Free Speech Is a Triangle" (2018) 118 Colum L Rev 2011 at 2021.

17 Andrew Tutt, "The New Speech" (2014) 41 Hastings Const LQ 235 at 240.

18 Andrew Tutt, "The New Speech" (2014) 41 Hastings Const LQ 235 at 243–246; Moran Yemini, "The New Irony of Free Speech" (2018) 20 Colum Sci & Tech L Rev 119 at 165–168.

on platforms, as well as navigation processes that direct users to certain content.[19]

16 This private governance, from the perspective of online communication systems, is a wide scope of power wielded by platforms. As Klonick argues, this power has resulted in a revolution in the infrastructure of free expression and changed the communication system in the digital era.[20] Since the problem of disinformation on platforms is also about the communication system, we need to focus on the power that affects online user speech and information circulation to identify the source of disinformation within the communication system.

B. Gatekeeping power and disinformation

17 This section will further clarify that from a communication system's perspective, the private governance of platforms is the gatekeeping power to control or even manipulate information to influence the production, expression and dissemination of user speech. Given that social media platforms have come to occupy an important structural position, it is argued that platforms also constitute part of the disinformation problem.

(1) The gatekeeping power of platforms

18 In general, gatekeepers are entities which decide what shall or shall not pass through. Gatekeepers can prevent misconduct because they control access to the tools, space, or community required to commit misconduct.[21] RH Kraakman used the term "gatekeeper liability" to describe the liability of gatekeepers, which is imposed by the public sector on private parties who are able to disrupt misconduct by withholding their co-operation from wrongdoers. According to Kraakman, the gate is crucial to wrongdoing.[22]

19 When it comes to social media, gatekeeping power is largely about curating online access to information and determining what content can be disseminated to users. The power to control information is in the

19 Sofia Grafanaki, "Platforms, the First Amendment and Online Speech Regulating the Filters" (2018) 39 Pace L Rev 111 at 117–118.

20 Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harv L Rev 1598 at 1663.

21 Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015) at p 37.

22 Such support are usually in the forms of specialised goods, services, or certifications. See Reinier H Kraakman, "Gatekeepers: The Anatomy of Third-Party Enforcement Strategy" (1986) 2 J L Econ & Org 53 at 54.

hands of platforms. They exercise this power to stifle the dissemination of illegal and harmful content to maintain profitability.[23]

20 There are two groups of gatekeepers, the first has direct control over information access, while the second, playing a facilitating role, controls access to the services needed to connect users to various content. An example of the former is the editor who decides what content is to be published, while operators of cable network channels are of the second type, given their facilitating role in information flow.[24]

21 This distinction is crucial in terms of the different influences on media diversity associated with different gatekeepers, but it is enough to point out that social media platforms may fall into both groups at the same time.

22 For example, platform algorithms control access to information when they inevitably block or delete content. Platforms also facilitate content, such as curating and disseminating content they think users would be attracted to.[25]

23 In short, gatekeeping power in the social media context is not merely the power to disrupt misconduct, but also the power to control the flow of information and even shape the whole communication system.

24 For example, in 2016, in response to the accusation of suppressing conservative news on its platform and its inability to prevent fake news, Facebook chose to change its algorithm to prioritise content posted by users' friends and relatives over news posted by traditional media. This change caused many content producers who relied on Facebook to reach their audiences to suddenly face a situation in which their audiences had been greatly reduced.[26]

---

23 Philip Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age* (Columbia University Press, 2019) at p 121.

24 Natali Helberger, Katharina Kleinen-von Königslöw & Rob van der Noll, "Regulating the New Information Intermediaries as Gatekeepers of Information Diversity" (2015) 17 *Info* 50 at 53–54.

25 Philip Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age* (Columbia University Press, 2019) at pp 122–123; András Koltay, *New Media and Freedom of Expression Rethinking the Constitutional Foundations of the Public Sphere* (Bloomsbury Publishing, 2019) at pp 84–85.

26 Philip Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age* (Columbia University Press, 2019) pp 69–70.

(2) Gatekeeping power and disinformation

25 The power that controls the flow of information and curates content for users nevertheless facilitates disinformation. Two cases are discussed to support this argument; one is the filter bubble phenomenon and the other is the result of micro-targeting technology.

26 The goal of platforms is not to provide the most objective or useful information, but to provide personalised information most relevant and most attractive to the individual.[27] However, in the long run, the pursuit of relevance leads to the filter bubble phenomenon, such that social media algorithms exclude people from information that do not correspond to their preferences or political orientations.[28] This is because on one hand, users are more likely to notice and click on information they prefer. On the other hand, opposing views are filtered out by platforms, because those views are deemed by platforms as irrelevant to users. In a filter bubble, users only experience views that echo their own – they are less likely to be exposed to opposing views.

27 Disinformation can be enlarged by filter bubbles: if all information in the filter bubble is false, it will be difficult for users to detect their own mistakes. Instead, they will constantly interact with friends and sources of information they see, which strengthens their own beliefs.

28 The filter bubble phenomenon exists when successful platform gatekeeping means controlling information to monetise, not rooting out disinformation. As long as people prefer to interact with content with questionable veracity, the filter bubble will deflect legitimate information that debunks the false information that had been previously consumed.[29]

29 The second case concerns micro-targeting technology. Platforms nowadays have unprecedented capacity to microtarget users. The more advanced the algorithms, the greater the gatekeeping power. The capacity to microtarget users allows everyone – both platforms and other users – to distribute certain information to specific audiences more accurately.

---

27 Brittainy Cavender, "The Personalization Puzzle" (2017) 10 Wash U Jur Rev 97 at 107; Siva Vaidhyanathan, *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy* (Oxford University Press, 2018) at 90.

28 Siva Vaidhyanathan, *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy* (Oxford University Press, 2018) at pp 91–92; Philip Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age* (Columbia University Press, 2019) at p 98.

29 Philip Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age* (Columbia University Press, 2019) at pp 98–99; Brittainy Cavender, "The Personalization Puzzle" (2017) 10 Wash U Jur Rev 97 at 108.

As Philip Napoli argues, targeting exclusively right or left-leaning news consumers has never been easier, as social media platforms have collected and processed large amounts of user data, which provides reliable indicators of an individual's political orientation. In these ways, false speech can achieve amplified results with unprecedented efficiency.[30]

30 Philip Howard and Samuel Woolley use the term "computational propaganda" to describe the phenomenon of digital manipulation and false information, which includes the use of algorithms, automation, and human curation to purposefully manage and distribute misleading information over social media networks.[31] As studies of computational propaganda show, it is possible to use algorithms to carry out political attacks, distribute disinformation, and create fake discussions, thereby manipulating public opinion and even producing a "manufactured consensus".[32]

31 Also, it was widely reported that in the 2016 US president election, the Trump campaign employed Cambridge Analytica, which drew upon massive amounts of social media data to construct detailed psychological, demographic and geographic profiles of individual voters, then used this data to deliver microtargeted political messages through social media platforms. This example shows that the technological capacity to target citizens with tailored messages or information based on their unique characteristics appears to be more advanced.[33]

32 Some studies have proven that disinformation can be effectively targeted at individuals who are more susceptible to false information. It can even be curated to specific targets. Those with economic, political or any nefarious agendas are now far better equipped to reach their target

30 Philip Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age* (Columbia University Press, 2019) at pp 96–97.

31 Samuel Woolley & Philip Howard, "Introduction: Computational Propaganda Worldwide" in *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media* (Samuel C Woolley & Philip N Howard eds) (Oxford University Press, 2019) at p 4; Samuel Woolley, "Bots and Computational Propaganda: Automation for Communication and Control" in *Social Media and Democracy The State of the Field, Prospects for Reform* (Nathaniel Persily & Joshua A Tucker eds) (Cambridge University Press, 2020) pp 98–100.

32 Samuel Woolley & Philip Howard, "Introduction: Computational Propaganda Worldwide" in *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media* (Samuel C Woolley & Philip N Howard eds) (Oxford University Press, 2019) at pp 4–5; Philip Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age* (Columbia University Press, 2019) at p 96.

33 Philip Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age* (Columbia University Press, 2019) at pp 96–98.

audience. If this capacity to microtarget users is combined with the filter bubble effect, disinformation on social media platforms will be spread even further.

33 To conclude, gatekeeping powers can amplify disinformation and lead to negative outcomes on the communication system through platforms. From the government's perspective, given that disinformation cannot be regulated effectively without managing the gatekeepers of information flow, platforms should be held partially accountable for the rapid dissemination of disinformation.

III. The flaws of self-regulation and meta-regulation

A. The self-regulation of disinformation

34 This Part first discusses some existing self-regulatory efforts that tackle disinformation. Secondly, it is argued that when it comes to addictive design, platforms cannot regulate disinformation effectively because they tend to practice reluctant regulation. While platforms may have enough ways to regulate disinformation, they might overly regulate the content they define as disinformation at the cost of users' freedom of speech, for they are not accountable to users.

(1) Self-regulating disinformation by platforms

35 Theoretically, platforms have existing self-regulatory mechanisms to deal with disinformation. Facebook and Instagram are working with independent, third-party fact-checking organisations to identify, review and take action on problematic communication shared on their platforms.[34] Not to mention since 2016, many platforms have already enforced mechanisms to combat the circulation of fake news.[35]

---

34 Ingrid Volkmer, *Social Media and Covid-19: A Global Study of Digital Crisis Interaction Among Gen Z and Millennials* (Wunderman Thompson, 1 December 2021) at p 67 <https://www.who.int/news-room/feature-stories/detail/social-media-covid-19-a-global-study-of-digital-crisis-interaction-among-gen-z-and-millennials> (accessed 8 December 2021).

35 Stephanie Ricker Schulte, "Fixing Fake News: Self-Regulation and Technological Solutionism" in *Fake News: Understanding Media and Misinformation in the Digital Age* (Melissa Zimdars & Kembrew Mcleod eds) (MIT Press, 2020) at pp 135–140; Alexandra Andorfer, "Spreading like Wildfire: Solutions for Abating the Fake News Problem on Social Media via Technology Controls and Government Regulation" (2018) 69 Hastings LJ 1409 at 1412–1413.

36 For example, Facebook has created an independent oversight board, giving said body a degree of authority to take down demonstrable falsehoods that might cause harm. Also, if information has been independently debunked by its third-party fact-checkers, the company will reduce its spread. Besides, content that has been rated false or partly false by a third-party fact-checker will be labelled to allow users to decide for themselves what to read, trust and share.[36]

37 According to Alexandra Andorfer, there are at least two ways to identify and regulate false content: human judgement and technological solutions.[37] In terms of human judgement, platforms can identify disinformation when users flag or report certain content. Moderators can then notify users what steps they can take to deal with flagged posts, such as removing or blocking. One strength of human judgement is that humans can assess content in their respective contexts to determine whether something is deliberately false.[38] As for technological solutions, automated fact-checkers use algorithms to check information and assess it against factual data. The benefit of using technological solutions is twofold: first, it tends to be more accurate than human judgement; second, artificial intelligence is less biased than humans.[39]

38 However, there are several concerns about algorithms. The first is the possible flaws in algorithm design. For example, the algorithm needs a programmer, but programmers may not take into account the integrity of the article or the integrity of the report when they design their systems. Their biases might be reflected in biased algorithms. The second concern of using algorithms to regulate speech is that platforms often do not publicise the criteria they use to identify false content, which

36 Cass Sunstein, *Liars, Falsehoods and Free Speech in an Age of Deception* (Oxford University Press, 2021) at pp 122–123.

37 Alexandra Andorfer, "Spreading like Wildfire: Solutions for Abating the Fake News Problem on Social Media via Technology Controls and Government Regulation" (2018) 69 Hastings LJ 1409 at 1413–1418.

38 Stephanie Ricker Schulte, "Fixing Fake News: Self-Regulation and Technological Solutionism" in *Fake News: Understanding Media and Misinformation in the Digital Age* (Melissa Zimdars & Kembrew Mcleod eds) (MIT Press, 2020) ch 10, at p 136; Alexandra Andorfer, "Spreading like Wildfire: Solutions for Abating the Fake News Problem on Social Media via Technology Controls and Government Regulation" (2018) 69 Hastings LJ 1409 at 1413–1415.

39 Alexandra Andorfer, "Spreading like Wildfire: Solutions for Abating the Fake News Problem on Social Media via Technology Controls and Government Regulation" (2018) 69 Hastings LJ 1409 at 1418–1419.

has a chilling effect, undermining the diversity of speech when it comes to controversial content.[40]

39 Given that platforms have ample experience in dealing with fake news, one might expect them to have the capacity to deal with disinformation. One might imagine that as long as platforms use one of the aforementioned methods, they should be able to take action against "disinformative" content. However, in the next section, it will be argued that leaving social media companies to regulate disinformation is more likely to end up in systemic failure: self-regulation may either end up in reluctant regulation or over-regulation.

(2) Addictive design and reluctant regulation

40 Platforms can be reluctant when it comes to regulation because they need to make profit. Reducing disinformation and maintaining a healthy information environment are not their main goals.[41] As some studies have shown, while platforms are technologically well-situated to minimise the amount of disinformation that may reach users, regulating disinformation might conflict with their business interests.[42] Content that generates more engagement and attracts more attention is not necessarily of better quality or veracity. Rather, it might just be more emotional, elicit more rage, fear, or other strong emotions, negative or not.[43] Unfortunately, "disinformative" content often gauges up user engagement. Thus, regulating disinformation might lead to reductions in profits.[44] Moreover, if regulating disinformation implies compromising

40 Alexandra Andorfer, "Spreading like Wildfire: Solutions for Abating the Fake News Problem on Social Media via Technology Controls and Government Regulation" (2018) 69 Hastings LJ 1409 at 1421.

41 Niva Elkin-Koren & Maayab Perel, "Guarding the Guardians: Content Moderation by Online Intermediaries and The Rule of Law" in *The Oxford Handbook of Online Intermediary Liability* (Giancarlo Frosio ed) (Oxford University Press, 2020) at p 671.

42 Abby K Wood & Ann M Ravel "Fool Me Once: Regulating Fake News and Other Online Advertising" (2018) 91 S Cal L Rev 1223 at 1245; Victor Pickard, "Confronting the Misinformation Society: Facebook's 'Fake News' Is a Symptom of Unaccountable Monopoly Power" in *Fake News: Understanding Media and Misinformation in the Digital Age* (Melissa Zimdars & Kembrew Mcleod eds) (MIT Press, 2020) pp 123–124.

43 Christian Stöcker, "How Facebook and Google Accidentally Created a Perfect Ecosystem for Targeted Disinformation" in *Multidisciplinary International Symposium on Disinformation in Open Online Media 2019* (Christian Grimme *et al* eds) (Springer, 2020) at p 142.

44 Abby K Wood & Ann M Ravel "Fool Me Once: Regulating Fake News and Other Online Advertising" (2018) 91 S Cal L Rev 1223 at 1246; Christian Stöcker, "How Facebook and Google Accidentally Created a Perfect Ecosystem for Targeted Disinformation" in *Multidisciplinary International Symposium on Disinformation*

(cont'd on the next page)

their business models to be less competitive, platforms will also be slow to take action, because they are obligated serve their shareholders.

41 Besides, when regulating disinformation brings a bad reputation or leads to controversies, platforms will also be reluctant to regulate.

42 As Cass Sunstein pointed out, although Facebook has enforced some methods to combat false content, there are several exceptions. For example, the exemption for politicians. Facebook does not want to touch on the issue of political ads even if there are clear and demonstrable errors. The reason is that politicians of all kinds would soon accuse their opponents of lying and ask Facebook to remove their ads. The decisions of the platform would predictably be subject to claims of political bias.[45] This will incur negative press. Moreover, some studies have shown that regulatory responses to disinformation have unsurprisingly become a partisan affair in the US. Social media companies are naturally eager to avoid accusations of partisanship.[46] As a result, platforms might well conclude that it is good for their businesses to adopt a general rule: allow a free-for-all. However, the effect of disinformation can instantly spread to countless people with the help of platforms via algorithms and personalisation.[47]

43 In short, when regulating disinformation is in conflict with business interest, platforms might be reluctant to regulate disinformation.

(3) Freedom of speech, due process and over-regulation

44 There are two dimensions of freedom of speech: substantive and procedural. In terms of substantivity, online platforms, unlike public actors, are not required to ensure the same constitutional safeguards when they make decisions over the organisation or removal of online speech. For example, in the US, while content-based regulation enforced by state actors would be prohibited under the First Amendment, constitutional law permits platform owners to engage in content-based regulation because they are private actors. Besides, platforms enforce speech norms

in Open Online Media 2019 (Christian Grimme *et al* eds) (Springer, 2020) at pp 143–145.

45 Cass Sunstein, *Liars, Falsehoods and Free Speech in an Age of Deception* (Oxford University Press, 2021) at pp 124–125.

46 Chris Marsden, Ian Brown & Michael Veale, "Responding to Disinformation: Ten Recommendations for Regulatory Action and Forbearance" in *Regulating Big Tech: Policy Responses to Digital Dominance* (Martin Moore & Damian Tambini eds) (Oxford University Press, 2022) at p 199.

47 Cass Sunstein, *Liars, Falsehoods and Free Speech in an Age of Deception* (Oxford University Press, 2021) at p 125.

that protect far less expression than the corresponding obligations of the government under the First Amendment. Platforms police abusive speech, sexual expression and hate speech, which might be shielded from government regulation by the First Amendment.[48]

45 On the other hand, there are also concerns about censorship. For example, in a country where denying the Holocaust is illegal, the government may put pressure on social media companies to police such speech. When social media platforms try to avoid conflicts with the government, they may end up overblocking speech that resembles, but is not, a denial of the Holocaust. For example, during the municipal election in Turkey in 2014, the Turkish Government blocked YouTube and Twitter for failing to comply with its national security laws to remove certain videos during the election.[49] This not only put pressure on social media platforms to delete content, but also made it difficult for Turkish people to discuss public affairs online.[50]

46 In terms of substantivity, since platforms are private actors, when they regulate online speech, they are not obligated to follow constitutional principles of free speech.[51] Thus, when platforms regulate disinformation, either by enforcing content moderation policies (by deleting or blocking disinformation) or by means of navigation processes (by sorting content with algorithms to reduce the number of reachable users), platforms can define disinformation as broadly as possible to their own convenience and enforce measures normally unavailable to the public sector.

47 In terms of procedures, because platforms are not accountable to their users, they do not govern their users in the same way liberal democracies govern their people.[52] The basic procedural obligations of those who govern populations in democratic societies include: (a) obligations of transparency, notice and fair procedures; (b) the offer of reasoned explanations for decisions or changes of policy; (c) the ability

48 Jack M Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1194.

49 Constanze Letsch & Dominic Rushe, "Turkey Blocks YouTube Amid 'National Security' Concerns" (28 March 2014) <https://www.theguardian.com/world/2014/mar/27/google-youtube-ban-turkey-erdogan> (accessed 15 December 2021).

50 Monika Bickert, "Defining the Boundaries of Free Speech on Social Media" in *The Free Speech Century* (Lee C Bollinger & Geoffery R Stone eds) (Oxford University Press, 2019) at p 258.

51 Jack M Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1195; Niva Elkin-Koren & Maayab Perel, "Guarding the Guardians: Content Moderation by Online Intermediaries and The Rule of Law" in *The Oxford Handbook of Online Intermediary Liability* (Giancarlo Frosio ed) (Oxford University Press, 2020) at pp 672–673.

52 Jack M Balkin, "Free Speech Is a Triangle" (2018) 118 Colum L Rev 2011 at 2035.

of users to complain about the conduct of the institution and demand reforms; and (d) the ability of users to participate, even in the most limited ways, in the governance of the institution.[53]

48 Platforms may often fail to meet these procedural and participatory fairness requirements, since private enforcement often lacks notice, due process and transparency. Though platforms usually claim that they exercise power benevolently and appropriately, there still can be arbitrary exceptions under their governance.[54] As Balkin pointed out, private governors reserve the right to act arbitrarily on occasion, and are much like 19th century enlightened despots:[55]

> They champion a set of enlightened values that they believe that their end-users want—or should want—but they implement these values through bureaucracy and code without taking any sort of vote.

That is to say, although users may accept, to some degree, that companies will follow internal policies to take down content, they may still be dissatisfied with the fact that policing criteria are kept hidden and rules are applied arbitrarily.[56]

49 As a result, when platforms regulate themselves, they tend to not give users due process, failing to provide transparency, notices and fair procedures. Platforms often do not offer users reasoned explanations for decisions or changes in policy, and users usually cannot participate in arbitration. In effect, platforms can regulate as much as they want to.

50 To conclude, self-regulation may either be ineffective or too effective. When platforms practise reluctant regulation, they tend to permit the circulation of "disinformative" content. When platforms practice over-regulation, companies might also overly restrict speech and thus negatively impact the online speech system.[57]

---

53 Jack M Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1197–1198.

54 Jack M Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1197; Barrie Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation" (2020) 43 Fordham Int'L LJ 939 at 959; Kyle Langvardt, "Regulating Online Content Moderation" (2018) 106 Geo LJ 1353 at 1385–1386.

55 Jack M Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1200.

56 Jack M Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1197.

57 Niva Elkin-Koren & Maayab Perel, "Guarding the Guardians: Content Moderation by Online Intermediaries and The Rule of Law" in *The Oxford Handbook of Online Intermediary Liability* (Giancarlo Frosio ed) (Oxford University Press,

(cont'd on the next page)

B. Meta-regulatory framework

(1) Why we need a meta-regulatory framework

51 The ineffectiveness of platform self-regulation demonstrates two points. First, in terms of gatekeeping, the self-regulation of platforms might not be successful because they lack the incentives to regulate disinformation properly. On one hand, although platforms have the resources and techniques, their gatekeeping performance will be impeded by their private interests, causing them to practise reluctant regulation. On the other hand, platforms might over-regulate disinformation without taking care of users' freedom of speech (both substantial and procedural dimensions) because they are not accountable to users.

52 Secondly, while platform self-regulation might be ineffective, the capacity of the government to directly enforce the regulation of disinformation on platforms is even more limited when compared to that of private platforms. Colin Scott used the metaphor of regulatory space to illustrate this point: in order to regulate a "space", the regulator must have resources. These resources are not limited to formal public authority, but also include regulatory knowledge, financial resources and organisational capabilities.[58] If we apply this metaphor to social media platforms, it is easy to point out that private platforms have all the resources necessary to enforce community norms. Generally speaking, platforms are more advanced than governments on technical grounds, so they can incorporate their domain knowledge in designing tools to regulate speech.[59] Also, companies interact with users more directly on their "spaces" and can modify their platforms anytime through user feedback.[60] In this way, the government is no longer the main *locus* of regulatory power over platforms.

53 More importantly, public policy debates about online platforms are highly dependent on data provided by private sectors. At present, policymakers not only do not know how to resolve the policy issues at

---

2020) at pp 670–674; Emily B Laidlaw, "Myth or Promise? The Corporate Social Responsibilities of Online Service Providers for Human Rights" in *The Responsibilities of Online Service Provider* (Maria Rosaria & Taddeo Luciano Floridi eds) (Springer, 2017) at p 151.

58 Colin Scott, "Analysing Regulatory Space: Fragmented Resources and Institutional Design" (2001) *Public Law* 283 at 284.

59 Colin Scott, "Analysing Regulatory Space: Fragmented Resources and Institutional Design" (2001) *Public Law* 283 at 297; Jack M Balkin, "The First Amendment in the Second Gilded Age" (2018) 66 Buff L Rev 979 at 998–999.

60 Colin Scott, "Standard-Setting in Regulatory Regimes" in *The Oxford Handbook of Regulation* (Robert Baldwin, Martin Cave & Martin Lodge eds) (Oxford University Press, 2010) pp 112–114.

hand, but also are unable to gain comprehensive information. That is, if governments want to enforce regulations, they need to have access to platform data. However, data released from private companies' voluntary provision might be problematic not just because the data is unverified, but also because they can selectively release data to influence policymaking.

54 Given the fact that the necessary resources for regulating social media are primarily in the hands of platforms instead of regulators, and that self-regulation can be ineffective, a crucial question emerges: what can governments do when the self-regulation of the regulated fails to meet government policy objectives? It is important to note that the goal is not to destroy current platform profiting models for the sake of regulating disinformation. The main task is to strike a fair balance between governmental regulatory needs and private interest. Such a balance merely indicates that the profit model of platforms should be modified to comply with public policy considerations.

55 To tackle disinformation through regulating platforms and to avoid over-regulation, governments need to establish effective enforcement of platforms' self-regulatory efforts and establish some procedural requirements. These requirements, provided by legal or liability structures, can create strong incentives for platforms to establish effective processes both for identifying and removing "disinformative" content, and for providing procedural protection to users.

56 The following section proposes the adoption of a meta-regulatory framework as a viable policy solution to help establish relationships between governments and regulated platforms, which will require platforms to achieve their specific objectives as well as maintain the advantages of self-regulation.[61]

(2) Why meta-regulation is a solution

57 Meta-regulation happens when outside regulators deliberately – rather than unintentionally – induce targets to develop their own internal, self-regulatory responses to public problems. It takes insight from self-regulatory approaches by which the regulated can be the source of their own constraint.[62] Under a meta-regulatory framework, the government

61 See Colin Scott, "Reflexive Governance, Regulation and Meta Regulation: Control or Learning" in *Reflexive Governance: Redefining the Public Interest in a Pluralistic World* (Olivier De Schutter & Jacoques Lenoble eds) (Bloomsbury Publishing 2010) at pp 62–63.

62 Cary Coglianese & Evan Mendelson, "Meta-Regulation and Self-Regulation" in *The Oxford Handbook of Regulation* (Robert Baldwin, Martin Cave & Martin Lodge eds) (Oxford University Press, 2010) at p 151; Colin Scott, "Reflexive Governance,

(cont'd on the next page)

would identify problems and require or force the regulated (private actors) to develop plans to solve them. The regulated would subsequently respond to the requirements by developing their own internal regulating system.[63]

58 Under this framework, the government can shape how the regulated regulates itself in various ways. The location and role of the government can vary. At one extreme, the government can be an active director who commands private actors by law to assist in the regulatory process. For example, in some Western nations, banks have been transformed into agents of the State and have become instruments of policy. They are required by law to routinely report transactions over a certain threshold and those transactions that are of a suspicious nature to a governmental authority.[64]

59 Less coercively, the government can provide rewards and incentives to induce the compliance of a regulated entity.[65]

60 Moreover, the government can be a facilitator or monitor of corporate social control exercised by non-governmental institutions. The government can "steer" rather than "row," structuring the marketplace to facilitate naturally-occurring private activity to assist in furthering public policy objectives.

61 In general, meta-regulation is typically characterised by its management-based commands, for regulators recognise their own limitations and explicitly encourage self-regulatory efforts. Instead of functioning as primary regulators, governments can oversee how targets self-regulate and whether they meet policy goals. Thus, the target will respond to public goals by developing what can be viewed as a self-

Regulation and Meta Regulation: Control or Learning" in *Reflexive Governance: Redefining the Public Interest in a Pluralistic World* (Olivier De Schutter & Jacoques Lenoble eds) (Bloomsbury Publishing, 2010) at pp 43–44.

63 Cary Coglianese & Evan Mendelson, "Meta-Regulation and Self-Regulation" in *The Oxford Handbook of Regulation* (Robert Baldwin, Martin Cave & Martin Lodge eds) (Oxford University Press, 2010) at p 150.

64 Peter Grabosky, "Meta-Regulation" in *Regulatory Theory: Foundations and Applications* (Peter Drahos ed) (ANU Press, 2017) at p 152.

65 Peter Grabosky, "Meta-Regulation" in *Regulatory Theory: Foundations and Applications* (Peter Drahos ed) (ANU Press, 2017) at p 152; Cary Coglianese & Evan Mendelson, "Meta-Regulation and Self-Regulation" in *The Oxford Handbook of Regulation* (Robert Baldwin, Martin Cave & Martin Lodge eds) (Oxford University Press, 2010) at pp 150–151.

regulatory system.[66] Meta-regulation can also be referred to as "enforced self-regulation", which requires private companies to bear the costs of establishing and enforcing their own systems.[67] Both meta-regulation and enforced self-regulation models require public authorities to set goals, monitor enforcement and interact with targets, but meta-regulation can provide a higher degree of discretion to targets and thus can be more flexible.[68]

62 Meta-regulation is marked by the degree of discretion it provides to targets. Given that targets have more knowledge and resources, they are more likely to solve the problems in efficient ways. Compared to traditional regulatory methods, which often directly require targets to take specific measures, thus restricting and depriving the discretion of targets, the meta-regulatory model delegates regulatory authority to targets and allows them to retain discretion despite policy requirements.[69] Under the meta-regulatory framework, platforms will integrate policy objectives into their present system and modify the system at a minimal cost. One can thus say that platforms play a role in balancing regulation enforcement and private interests, and the balance is struck not by strict public authority but by platforms themselves.

63 Therefore, the inability of the public sector to regulate can be alleviated to some degree because the public sector becomes a monitor and not a direct regulator. The main task is to facilitate or monitor platforms to use their available or potential technological solutions and identify the most efficient regulatory measure to regulate disinformation properly. In short, the meta-regulatory model takes advantage of target information and resources to make regulation effective.[70]

---

66 Cary Coglianese & Evan Mendelson, "Meta-Regulation and Self-Regulation" in *The Oxford Handbook of Regulation* (Robert Baldwin, Martin Cave & Martin Lodge eds) (Oxford University Press, 2010) at pp 150–151.

67 John Braithwaite, "Enforced Self-Regulation: A New Strategy for Corporate Crime Control" (1982) 80 Mich L Rev 1466 at 1470–1471.

68 See Colin Scott, "Reflexive Governance, Regulation and Meta Regulation: Control or Learning" in *Reflexive Governance: Redefining the Public Interest in a Pluralistic World* (Olivier De Schutter & Jacoques Lenoble eds) (Bloomsbury Publishing 2010) at pp 61–63.

69 Cary Coglianese & Evan Mendelson, "Meta-Regulation and Self-Regulation" in *The Oxford Handbook of Regulation* (Robert Baldwin, Martin Cave & Martin Lodge eds) (Oxford University Press, 2010) at pp 151–152.

70 Cary Coglianese & Evan Mendelson, "Meta-Regulation and Self-Regulation" in *The Oxford Handbook of Regulation* (Robert Baldwin, Martin Cave & Martin Lodge eds) (Oxford University Press, 2010) at pp 153–154.

(3) Why meta-regulation is justified

64 This article argues that meta-regulation can be justified in terms of the gatekeeping power of platforms. This power, which sets and enforces the rules that manage and process data, has allowed platforms to control the flow of information for business purposes. This power has also led to an increase in the economic and political power of some private actors in the digital age, where monopoly over information no longer belongs exclusively to public authorities, but also to private actors.[71]

65 Besides, this gatekeeping power also defines the standard of protection of fundamental online rights. While from a constitutional law perspective, this power has traditionally been vested in public authorities, the gatekeeping power of private platforms has shaped standards of protection and procedures, further determining who can say what, what can be seen, and what should be deleted; and thus impacts users' freedom of speech online.[72]

66 The negative consequences of disinformation are amplified by how platforms gate keep. Given the fact that self-regulation is not effective, governments have legitimate reasons to regulate platforms in tackling disinformation, by regulating how platforms self-regulate.

67 On the other hand, imbalances between users and platforms, such as the lack of accountability and transparency safeguards, justify the policy objective of protecting users' freedom of speech. Since information is organised by business interests, driven by profit rather than democracy, transparency or accountability, it is necessary to focus on the positive dimensions of public authorities to introduce procedural safeguards of free speech.[73]

68 Meta-regulation can be seen as a response to the exercise of gatekeeping power of platforms. The goal is to ensure the implementation of public policies online (regulating disinformation) while protecting the fundamental rights of users and avoiding violating the interests of platforms.

---

71 Giovanni De Gregorio, *Digital Constitutionalism in Europe Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press, 2022) at p 18.

72 Giovanni De Gregorio, *Digital Constitutionalism in Europe Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press, 2022) at p 29.

73 Giovanni De Gregorio, *Digital Constitutionalism in Europe Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press, 2022) at p 36.

IV. Meta-regulation and its concerns

A. Applying meta-regulation to tackle disinformation

(1) Platforms as governors

69 Under the meta-regulatory model, the regulator has two policy objectives: first, to enforce the regulation of disinformation; second, to require platforms to protect user speech in both substantive and procedural dimensions when they regulate disinformation. Social media platforms should develop internal self-regulatory plans to achieve these two objectives.

70 As discussed above, platforms have more regulatory resources in the regulatory space of social media than public sectors. Take user speech as an example; though governments can illegalise certain types of speech, only platforms can enforce regulations directly. Besides, compared to governments, platforms can more easily locate the place and time of such speech, and are better positioned to take action, such as by reducing reachable users, deleting or tagging.

71 By setting policy objectives and enforcing self-regulation, meta-regulation allows platforms to maintain the discretion of choosing their own methods. In the following section, this article will discuss two policy objectives respectively and lay out a meta-regulatory model for platform disinformation.

(2) Policy objective 1: Enforcing the regulation of disinformation

72 In terms of regulating disinformation, the government should set policy objectives to illegalise disinformation and enforce rules that reduce disinformation. It is important for governments to require social media platforms to play a more active role in enforcing regulation, because the intervention of public authority steers platforms away from reluctant regulation.

73 The government can explicitly obligate social media platforms to enforce the regulation of disinformation with or without stipulating concrete methods, giving them discretion to choose their preferred methods.[74] They have different ways, such as fining, to ensure that their requirements are met.

---

74 Giancarlo Frosio & Martin Husovec, "Accountability and Responsibility of Online Intermediaries" in *The Oxford Handbook of Online Intermediary Liability* (Giancarlo Frosio ed) (Oxford University Press, 2020) at p 621.

74 If platforms allow illegal content such as disinformation to circulate, they will be held responsible for the illegal content. In this regard, the onus of blocking illegal content is on platforms, because any delays in removing illegal content can enlarge the harm. The key is to reduce the harm done by "disinformative" content, rather than identify and punish people who post it. Under meta-regulation, platforms do not have formal public authority to punish or criminalise users; instead, platforms are obligated by law to enforce rules that aim at curbing the spread of disinformation.

75 In Germany, there is a law called the Network Enforcement Act ("NetzDG"). One of the obligations it imposes on platforms is that platforms must delete obviously unlawful content within 24 hours. It is important to note that "unlawful content" is not defined in this act, but in other German criminal laws. The NetzDG enforces regulation and protects users against unlawful content rather than prosecute people, and also incentivises platforms to remove unlawful content faster.[75] From a meta-regulatory perspective, NetzDG provides a possible way to enforce the regulation of disinformation. In which case, platforms are considered an instrument for public actors to ensure the enforcement of public policies.

76 By requiring platforms to enforce part of the regulations, meta-regulation can be seen as similar to the "new-school" speech regulation.[76] According to Balkin, new speech regulation often aims at regulating information intermediaries, through which users convey or receive information. One substantial feature of its "newness" is the regulatory power aiming at one thing to control another, such as aiming at social media to supervise user speech. Balkin argued that in the digital era, nation states often put pressure on intermediaries to enforce *ex ante* methods, including filtering and blocking, or *ex post* methods, such as takedowns with or without notice, to regulate illegal speech on platforms.[77]

77 The difference between meta-regulation and Balkin's new speech regulation is in the discretion of platforms, which can be strict or loose, depending on the design. In looser versions, the government does not need to require platforms to adopt specific methods to deal with

---

75 Amélie Heldt Pia, "Reading Between the Lines and the Numbers: An Analysis of the First NetzDG Report" (2019) 8 *Internet Policy Review* 1 at 2.

76 Jack M Balkin, "Old-School/New-School Speech Regulation" (2014) 127 Harv L Rev 2296 at 2306; Jack M Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1174.

77 Jack M Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation" (2018) 51 UCD L Rev 1149 at 1176–1177.

disinformation. The government can only require platforms to reach the public policy of "dealing with disinformation". The exact methods and details are left to platforms to decide. As for the stricter versions of meta-regulation, the government can explicitly specify what methods platforms should use. For example, the government can explicitly state under law that platforms must censor, silence, block, hinder, delay or delink false posts in a certain period of time. However, the stricter the version it is, the easier the government can abuse its powers when regulating platforms and users. The following section discusses the concerns about the risks of meta-regulation.

(3) Policy objective 2: Provide substantive and procedural protection while regulating disinformation

78 When regulating disinformation, it is equally important for platforms not to over-regulate users' speech. The second policy goal of meta-regulation is avoiding over-regulation. As discussed above, over-regulation might happen in either substantive or procedural dimensions. The NetzDG mentioned above also requires platforms to provide more clarity on the way platforms handle and moderate unlawful content. The goal of this requirement is to prevent platforms from over-regulating.

79 This section proposes an abstract model that provides a framework for flexible designs. Based on Emily Laidlaw's "Governance Model", this model focuses on the interactive relationship between public sectors and private corporations as well as the discretion that private corporations have.[78] While Laidlaw aims to provide a framework to protect all the rights of internet users, and her framework requires platforms to modify their business practices to a greater extent, this article only applies this model as far as disinformation regulation is concerned.[79]

80 Laidlaw's Governance Model has three layers: (a) education, research and policy; (b) government co-operation with companies in respect of policy formation, assessment, and auditing and advisory services; and (c) a remedial mechanism, including rule-setting and adjudication (see Figure 1 below).[80] Laidlaw states that "what sets this model apart from other frameworks is the responsive nature of the

78 Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015) at p 249.

79 This article narrows down the application of Laidlaw's model to enforcement of regulating disinformation, so the model can be more easily justified in its more limited scope.

80 This original figure is "Figure 6.3 Internet rights governance model" by Emily Laidlaw. *See* Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015) at p 259.

interplay between these layers of regulation, and the aim of this interplay is to facilitate integration of human rights within a business's operations".[81] The goal and the responsive nature of her model are similar to those of meta-regulation, for they aim at integrating the policy goals of protecting user speech while regulating disinformation within data-driven business models, and highlights the relationship between public and private sectors.

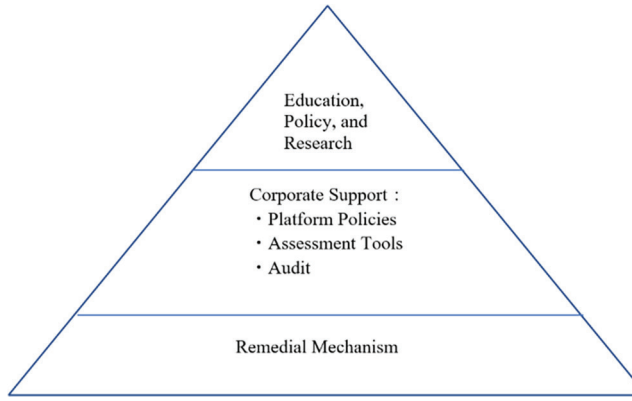

Figure 1: The government model proposed by Laidlaw

81 The base layer of the model is the remedial mechanism, which includes the internal adjudication of each case and the rule-making arm of the government. In order to deal with the lack of transparency and due process in platform governance, the remedial mechanism needs to allow users to access the adjudication process, as well as establish a consistent and reliable remedial mechanism for rule-setting. However, according to the meta-regulatory model, the design of these layers would depend on what is particular to each platform, and the operation would also be subject to different feasibility constraints. Therefore, there is no universal solution for remedial mechanisms. That said, there are still recommendations for best practices.

82 First, given that the greatest challenge of private governance is the lack of transparency to their users, rulemaking must be transparent.[82]

81 Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015) at p 259.

82 Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015) at p 263; Barrie Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation" (2020) 43 Fordham Int'l LJ 939 at 959.

In recent years, a number of proposals have been put forward by scholars and civil society groups, stating that platforms should ensure more structured multi-stakeholder participation in the development and revision of content moderation rules. Online platforms have taken certain steps. For example, Facebook has stated that they regularly invite external experts to join their content policy team meeting every few weeks to discuss its moderation policies. While these developments have resulted in some progress concerning the platform's moderation policies, they can still go further. To name a few: they can adopt notice-and-comment procedures to obtain public feedback on proposed changes to moderation policies, appoint outside experts in the form of an advisory panel to inform their policy decisions, or support the creation of independent multi-stakeholder bodies to help ensure the compatibility of platform moderation policies.[83]

83 Under a meta-regulatory framework, these proposals are subject to feasibility constraints depending on the specifics of each platform. The point is that platforms can also benefit from a more structured and sustainable approach to stakeholder engagement concerning their rule-making processes because users can participate in the rulemaking process to some extent.

84 Second, beyond rulemaking, platforms should also address transparency issues concerning their human and algorithmic decision-making processes. According to Barrie Sander, transparency has both quantitative and qualitative dimensions. The former refers to the statistical information disclosed by online platforms concerning their content moderation systems, and the latter is through independent forms of verification and auditing that aim to identify and assess the actual and potential impacts of these systems on users.[84]

85 After the first step of transparent decision-making, the second step is to ensure the standardisation and effectiveness of the process: users need to be notified when their posts are deleted or accounts are blocked as a result of disseminating disinformation; users also should be given opportunities to participate in the process,[85] which should not be

83 Barrie Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation" (2020) 43 Fordham Int'l LJ 939 at 991.

84 Barrie Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation" (2020) 43 Fordham Int'l LJ 939 at 992–993.

85 Giancarlo Frosio & Martin Husovec, "Accountability and Responsibility of Online Intermediaries" in *The Oxford Handbook of Online Intermediary Liability* (Giancarlo Frosio ed) (Oxford University Press, 2020) at p 620.

cumbersome, costly, or time-consuming, otherwise users will not find it helpful.

86 Moreover, the requirement of transparency is also important to the government, for public authorities can better understand how social media platforms identify and regulate disinformation and take further steps to "meta-regulate" platforms. For example, the government can require platforms to report what they do to regulate and how effective it is after a period of time to examine the outcomes of meta-regulation. On the one hand, such reports can show whether platforms provide users with due process and whether their procedures are sufficient to protect user rights. On the other hand, these reports can show whether platforms have done enough to deal with illegal content.

87 The second layer of the "Governance Model" is the most promising yet complicated layer, because it depends on the interaction between the meta-regulator and the targets under different conditions. The goal of this layer is to form coherent policies in the long run, so that platforms would be able to self-govern in accordance with public policies. Besides, independent audits for platform compliance can also prevent platforms from abusing their powers when it comes to regulation. In this layer, the public sector can act as a helpline to provide advice when companies need guidance in dealing with practical problems related to regulating disinformation.[86]

88 At the top of this pyramid is education, research and policy. In this layer, information is the key to the interaction between the public, the government, and platforms.[87] By requiring companies to provide an annual report of regulating disinformation and its impact on users to the government and the civil society, the public can receive useful information about what platforms have done to fulfil their responsibilities. Besides, platforms also need more information from the government and feedback from the public to optimise their policies. For example, while platforms are directly identifying what is disinformation, they need to follow the government's law that defines disinformation. This layer requires platforms to disclose their mechanisms and records to the public in order to ensure their regulatory efforts will not go beyond the law and thus unduly affect users' freedom of speech.

86 Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015) at p 268.

87 Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015) at p 260.

89 Through enforcing the regulation of disinformation and requiring platforms to protect user speech in both substantive and procedural dimensions, the meta-regulation of disinformation can establish an effective link between platforms and the government to resolve the shortcomings of self-regulation, while still allowing platforms some level of discretion. More importantly, this framework remains flexible and allows more interaction and wiggle room between platforms and the government in the future.

90 However, there are concerns and risks in applying meta-regulation. The last section will address the concerns about the increasingly asymmetrical power of platforms over their users. It will also discuss the risk of delegating public authority, which might lead to abuse of power by either governments or platforms.

B. Concerns of meta-regulation

(1) Overblocking and asymmetrical power over users

91 Considering how much gatekeeping power platforms have, some might argue that meta-regulation will give platforms excuses to abuse their power. Due to the fact that it is difficult for platforms to distinguish illegal disinformation from other legal content, when platforms are obligated to enforce regulation, they may remove questionable speech altogether in order to save operational costs. This is called overblocking.

92 When there is overblocking, content will be deleted or blocked even if there are no important reasons. This is because the incentive for immediate deletion is greater than following basic procedures. In this case, users may think that if they publish controversial content, they will be removed from the platform, contributing to a chilling effect.[88]

93 Overblocking is a real issue. Social media platforms already have the relevant technology to block and remove certain controversial content, and they also have the incentive to do so in order to create more profit.

94 Meta-regulatory efforts can be designed to restrict gatekeeping power and avoid overblocking by adhering to substantive and procedural protection while regulating disinformation. Users can have access to a more transparent process when their content is blocked. For the government, meta-regulation can help officials understand the actual

88 Amélie Heldt Pia, "Reading Between the Lines and the Numbers: An Analysis of the First NetzDG Report" (2019) 8 *Internet Policy Review* 1 at 4.

gatekeeping processes platforms exercise, not to mention provide access to the necessary resources and knowledge for more effective regulation.

(2) The risk of delegating public authority

95 The policy goal of regulating disinformation under meta-regulation is to officially authorise platforms the power to control speech. This shift to private governance might create the first risk of accountability concerning users' freedom of speech when private actors perform traditional public functions.[89]

96 Nevertheless, the online enforcement of public policies has been in the hands of platforms for a long time. They have enjoyed a broad margin of discretion in deciding how (or whether) to implement the law of each country. By virtue of governing their digital spaces, online platforms have performed autonomous quasi-public functions without the oversight of a public authority.[90] In this way, the primary issue for governments is not the delegation of power, but whether this delegation is legitimate and necessary. This article argues that such delegation is both necessary and legitimate. Considering legitimacy, one must realise the fact that social media platforms have become gatekeepers of information is irreversible. As discussed in Part IV.A, incorporating social media into regulatory frameworks can not only help enforce regulation of unlawful speech, but also prevent private companies from over-regulating speech without due process. Considering necessity, given that governments already have difficulties in regulating unlawful content on platforms, there are several ways out: (a) one is to fully rely on platform self-regulation; (b) the second is to establish a more detailed monitoring and tracking system; and (c) the third is to adopt a meta-regulatory framework.

97 In previous sections, this article argued that complete self-regulation is equivalent to letting go of enforcement, allowing unlawful content to spread rampantly on platforms.

98 Establishing a more detailed monitoring system will lead to the expansion of power by the State. One problem is that any overall monitoring and tracking system would come with immense costs. On the other hand, such system would provide governments the capacity for total surveillance, which fully interferes with the operation of platforms. Meta-regulation is the middle ground.

---

89 Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press, 2015) at p 40.

90 Giovanni De Gregorio, *Digital Constitutionalism in Europe Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press, 2022) p 97.

99 However, it is important to keep in mind that the power to meta-regulate can also be abused. Some might argue that meta-regulation is a double-edged sword. On the one hand, it is a more effective means for the government to achieve policy objectives; on the other hand, it may lead the government to abuse its power when it comes to enforcing regulation – governments may collude with companies to censor online speech, worsening the vulnerable state of users. Meta-regulation may also set objectives too harshly for platforms to achieve in order to punish or control platforms to secure government power. In these situations, either users or platforms will be vulnerable to the abuse of power.

100 Such concerns can be alleviated when specific meta-regulatory frameworks are designed under each State's constitution, which serves to limit the powers of the government. As the limits of power will vary in different countries, meta-regulation cannot be evaluated without analysing specific constitutions, which serve to prevent abuses of power.

V. Conclusion

101 This article has discussed the addictive design of social media platforms and how it effects private governance. It argued that from a communication system's perspective, platforms have the gatekeeping power to control or even manipulate information to influence the production, expression and dissemination of user speech. The exercise of such power also constitutes part of the disinformation problem. As a result, disinformation cannot be solved without taking platforms into consideration.

102 While platforms have taken certain efforts to regulate disinformation, such regulatory efforts tend to devolve into reluctant regulation or over-regulation.

103 Due to these drawbacks, this article applied meta-regulation to tackle disinformation. It is suggested that meta-regulation can effectively regulate disinformation without falling into the trap of over-regulation.

104 Finally, this article discussed concerns about the increasingly asymmetrical power of platforms over users and the risk of delegating public authority to private companies. Given that overblocking is an existing phenomenon that platforms tend to engage in, meta-regulation is needed to tackle such a difficulty. While it is possible for the government to abuse its power via meta-regulation, it is still necessary and legitimate to delegate some public authority to private platforms. In the end, each State's constitution should be the foundation for designing meta-

regulation to prevent governments from abusing their power through meta-regulatory efforts.