Harmful Speech Online: Five Models of Platform Regulation

Gavin Phillipson and Robert Mark Simpson

In Eric Heinze, Natalie Alkiviadou, Tom Herrenberg, Sejal Parmar, and Ioanna Tourkochoriti (Eds.), T*he Oxford Handbook of Hate Speech* (Oxford: Oxford University Press, 2026).

Abstract. How, if at all, should governments be involved in prescribing content moderation policies for extreme speech on social media platforms? There are a range of regulatory models that may be adopted, on a spectrum from *laissez-faire* self-regulation to punitive state control. We propose a normative taxonomy of such schemes, that is, a classification based on how the different regulatory models bear upon the normative legitimacy of state (in)action in this area. We defend a certain kind of hybrid regulatory model (i.e., government working with platforms on regulatory policy), which is legislative, rather than informal or advisory, so as to ensure democratic input, transparency, accountability and redress (due process), but which imposes procedural requirements upon companies' moderation policies and practices, rather than punitively prescribing moderation outcomes in specific cases.

1. Introduction

Should governments be regulating content moderation, removal, and demotion practices directed at harmful speech on social media platforms?[1] The US Constitutional answer, in essence, is "no". Generic worries about government interference with private actors apply with particular force to social media companies because the government

[1] For 'demotion' see text to n 28 below. By 'harmful speech' we mean to encompass a range of categories, including mis/disinformation, hate speech, advocacy of terrorism, threats, revenge porn, child sexual abuse material and so on. We are particularly interested in the patently harmful instances of these types, which are already unlawful in many countries.

cannot be allowed to proscribe or interfere with the content and viewpoint decisions they make. These companies should be left to self-regulate as they see fit.

This *laissez-faire* stance can seem irresponsible to European eyes. But there can also be something irresponsible about simplistic calls for urgent government intervention in this field. There's a compelling case, explored in what follows, that the significance of content moderation for democratic discourse is just too important for it to be governed by self-regulating corporations. However, you don't have to be a diehard Libertarian to see that state control over social media content can lead to authoritarian abuses, like election-fixing, or the suppression of political dissent under the guise of combatting 'terrorist content.'[2]

Debates about the justifiability of state-based platform regulation can easily reach an impasse bet ween these binary, opposing perspectives: state control *or* private self-regulation; Europe *vs* the US. We want to help move discussion forward by setting up a normative taxonomy of models for how governments might respond to the issue of platform regulation. By a *normative taxonomy*, we mean a classification that categorises the possible schemes based on the descriptive features relevant to their normative evaluation. Specifically, we focus on how different regulatory models are liable to affect government legitimacy in democratic states.[3] We will limit our discussion of laws and policies to those from the two regulatory superpowers, the US and the EU. But we won't be delving into the legal or institutional minutiae of any particular scheme.[4] Of course, the legitimacy of any specific regulatory regime, in a specific jurisdiction, will depend at least in part on these minutiae. What we are proposing here, though, is a model-based analysis: a more schematic way of assessing the legitimacy-related pros and cons of the various approaches to platform regulation. The details of existing policies will help to illustrate the distinctive features of different approaches. But at the evaluative stage we want to abstract away from such details. The goal is to develop a framework that can help NGOs, lawmakers, academics, and other stakeholders to evaluate the merits of different regulatory approaches, at the point where big-picture policy options – including those that prohibit any role for the state – are being considered, or where their ongoing utility is being questioned.[5]

Here then is a summary of our proposed taxonomy of five models of platform regulation:

[2] See below, 000-000.

[3] We confine our survey to broadly liberal-democratic states and legal orders (the EU) and so do not consider the far more authoritarian policies and practices of states like China.

[4] Nor does this paper purport to provide a comprehensive summary of all relevant EU regulation in this area, which is prolix and includes measures we do not further consider e.g., the specific regulation in relation to Video-Sharing Platforms: Directive 2010/13 of the European Parliament and Council of 10 March 2010 (Audio-Visual Media Services Directive), OJ [2010] L 95/1, as amended by Directive 2018/1808. For a good overview, albeit from 2020, see www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718 _EN.pdf.

[5] As is broadly the present position in the US: below at 000.

(A) ‘Pure’ Self-Regulation

‘Midway’ schemes

(B) Self-Regulation qualified by informal government influence

(C) Non-Legislated Hybrid Schemes

(D) Legislated Hybrid Schemes

(E) Punitive State Regulation

(By hybrid schemes, we mean ‘regulated self-regulation’ – the supplementing and reforming of self-regulation with elements of state intervention.[6]) As the headings indicate, the sequence (A) to (E) tracks a gradient from more *laissez-faire* schemes to greater government control. Small-government libertarians and traditional civil libertarians will tend to see model (A) as the best approach, and (E) as worst, all else being equal. And this corresponds with the US First Amendment approach in recent decades. We use the term *Statists* for people at the other end of the spectrum, whose main ethical concern is that democratic states should take responsibility for standard-setting that seeks to mitigate the impact of harmful speech online, and provide individuals with opportunities for redress. Statists will therefore tend to view model (A) as the worst approach. Models (C)–(E) correspond with current European approaches: both EU initiatives and hybrid regulatory schemes adopted by various European states.

We define a *Midway* scheme as one that answers, to some extent, to both Libertarian and Statist concerns. Advocates of these schemes see that Punitive State Regulation is liable to authoritarian abuse; but they also recognize that Pure Self-Regulation both lack democratic pedigree, and may be ill-equipped to redress the damage of harmful speech online, given that companies may lack a strong commercial incentive to carry out this work, or may be ideologically opposed to doing so.[7] They therefore tend to favour schemes like (B) (C), or (D), which eschew the extremes. An important aim of this paper is to highlight legitimacy-related distinctions between these ‘midway’ schemes.

In assessing these models, we will use five legitimacy-related criteria:

- Abusability
- Transparency
- Democratic Input
- Accountability, and
- Redress.[8]

[6] Below, n 76

[7] See below text to and n 24.

[8] As briefly explained in the Conclusion below, we do not in this paper attempt to use a sixth possible criterion, namely *efficacy*; this is partly because it would be practically impossible to make a proper assessment of the likely effectiveness of all the schemes we survey in a paper of this length, but also because our primary concern is the relative *legitimacy* of the broad models we survey rather than their (likely) effectiveness.

'Abusability' is a short-hand way of saying 'susceptibility to being abused or co-opted by authoritarian regimes'. So, this first criterion is about how resistant or open different models are to such abuse. The second criterion, *transparency*, is about how far the operations of a given content moderation scheme – especially any government influence or control over it – are rendered observable to citizen. The third criterion, democratic input, concerns the extent to which a given scheme has some kind of democratic imprimatur, via the intervention of norms or standards that have come about through some recognised process of democratic deliberation. The fourth criterion, *accountability*, pertains to legal accountability, i.e. the extent to which a scheme allows citizens and other parties to legally challenge government interventions in the content moderation practices of platforms. (Political accountability matters too, of course, but we are thinking of that as being captured under the transparency criterion.) Finally, the fifth criterion, *redress,* is about the ability of each model to provide remedial processes whereby people whose speech (or whole account) is demoted or removed from a platform, have effective ways of having that removal reviewed and potentially overturned.[9]

The first three criteria are primarily concerned with the democratic legitimacy, including contestability, of any given scheme and any governmental role in it. The last two are more focused on rule of law-type considerations. Our basic approach is to consider how the features of different models are liable to create legitimacy deficits, in respect of each of these criteria. So, we count the best model, roughly, as the one that is least susceptible to creating significant legitimacy deficits, across any of the criteria.

It will by now be apparent that we are working with a multi-dimensional conception of government legitimacy. For some readers, it may seem simpler to just view this exercise as an overall normative ranking, and to leave claims about legitimacy out of it – especially given the pervasive disagreement about exactly what government legitimacy involves, and how various factors bear on it. But we prefer to frame our analysis in terms of government legitimacy, as a way of consciously foregrounding a perspective that some work on the subject of platform regulation downplays. Given the immense power that citizens delegate to their governments, what can they expect from their governments, vis-à-vis their regulatory response to harmful speech online? Big tech companies are reshaping the world, for better and worse. We want to home in on the citizen-government relationship, in the face of this corporate and technological transformation. The citizen can, and must, ask of their government: what are you doing about this, and how?

[9] A common feature in reports by Human Rights Bodies and in civil society-authored sets of principles to govern this issue is calls for transparency, accountability and meaningful redress for those whose online speech is removed or restricted. See the Manila Principles (https://www.eff.org/files/2015/10/31/manila_principles_1.0.pdf, calling for strong due process, including remedial rights for users (Principle V), and for transparency and accountability to be built into law and content moderation practices (Principle VII); the 2018 Report on Online Content Moderation by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (https://www.ohchr.org/en/documents/thematic-reports/ahrc3835-report-special-rapporteur-promotion-andprotection-right-freedom) which calls for similar guarantees, and to similar effect, Council of Europe, *Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content regulation,* https://rm.coe.int/content-moderation-en/1680a2cc18 - see especially Part VII – 'Seven Key Principles for a Human Rights-Based Approach to Content Moderation'.

When engaging with this subject, it's important to note that the removal from platforms of various kinds of potentially harmful speech (and other content, such as copyright material) is going on all the time, and at enormous scale, regardless of what role the state has. To give an idea of this, in just three months in 2023 Facebook took down over 914 million pieces of content (over 10 million removals per day); in the same three months, YouTube removed 4.5 million videos, and TikTok over 102 million. Those numbers, huge as they are, are likely smaller than the times content is reviewed but left in place, (and sometimes demoted). So, our question is not whether this activity is a good or bad thing in terms of free speech. Rather we are addressing a more specific issue: what (if any) is the most legitimate role for the state in this? What can citizens require of their governments, in responding to the immense power of tech companies to organise, edit, and curate public discourse? We believe our taxonomy can help to structure debates between supporters of different models, clarifying some of the normative and empirical hinges on which they turn. For example, some commentary on this issue stresses the advantages of cooperative or 'hybrid regulatory approaches, while ignoring the ways in which different variants of such schemes differentially affect state legitimacy. Whatever the legitimacy-related upsides there may be in (as we will propose) eschewing models (A) and (E), this won't lead to more legitimate governance if those models are replaced with opaque regulatory schemes, or practices that preclude or obstruct democratic accountability or individual redress.

In particular we believe that our taxonomy can highlight some endemic problems of transparency, accountability, and due process that are liable to arise under models (B) and (C), in order to argue for Legislated Hybrid Schemes (D) as the preferable approach, other things being equal. We unpack this argument in §4. To set the stage, in §2, we explain the legitimacy-related downsides of a hard-line Libertarian or Statist approach, and then in §3 we outline the key descriptive features of the three Midway Schemes.

2. Libertarianism and Statism

2.1 Self-Regulation

Model (A), Pure Self-Regulation, is, formally speaking, the operative model in US law, where section 230 of *The Communications Decency Act* (CDA 230) gives social media platforms complete control over content moderation, and offers them a very broad shield from legal liability in respect of the content they host,[10] even when they *know* it to be unlawful. Thus, notifying a platform of unlawful content does *not*, as under the EU's long-standing qualified liability shield (p 00 below), trigger liability in the case of non-removal. Hence, in today's US, the main guardrails on what kind of speech can appear in online discourse are devised and implemented by the self-regulatory content moderation practices of private corporations. Distrust of government and its potential for over-

[10] Exceptions include material that breaches copyright, amounts to a federal crime, or to online sex trafficking. See also new legislation in the specific area of revenge porn: n 34 below.

reach and abuse allow private companies to occupy these immensely important 'gatekeeper roles', 'shap[ing] the public sphere' through 'their own "community standards"' based on their largely profit-driven interests – standards whose formulation has been dubbed a "quasi-legislative function"[11] in defining permitted speech on platforms.[12] These can range from the elaborate schemes seen on major platforms like Facebook, to the minimal-to-non-existent on fringe sites like 4chan. The point is that it is for platforms themselves to decide what, if any, content-based standards to impose.

The legitimacy-related advantages of Pure Self-Regulation (the *pure* here meaning simply the absence of a governmental role) are obvious.[13] Government is responsible, in the first instance, for its coercive acts, via law enforcement or its administrative agencies. When government refrains from regulating content moderation on platforms, it thereby refrains from carrying out coercive activity proscribing the content and especially the viewpoints of permitted speech. A strong and well-known school of free speech scholarship argues that such restraint is vital for the maintenance of governmental democratic legitimacy.[14] We accept a weaker version of this claim, namely that any time government involves itself in regulating the content of speech, even for reasonable, harm-prevention goals, it establishes precedents and mechanisms that could potentially be used to restrict speech in ways that are politically or ideologically persecutory. All speech regulation incurs this risk. Even critics (like ourselves) who are in favour of a more hands-on approach to speech regulation, have to acknowledge that a pro tanto consideration in favour of Pure Self-Regulation is that it is less susceptible to this kind of capture and abuse.

It is also important to guard against simplistic perceptions of self-regulation as resting purely on the 'internal' standards and processes of the platforms themselves. Platforms can and do reach out to civil society for expert assistance with content moderation. Particularly in relation to potentially harmful health-related disinformation, many social media companies, in the US as elsewhere, have sought input on how to implement their in-house speech guidelines from third-party fact-checking groups. Notably, during the Covid-19 pandemic, many platforms coordinated with the World Health Organisation in deciding which particular forms of pandemic misinformation to remove or demote.

[11] Belli, Luca, and Jamila Venturini (2016) "Private Ordering and the Rise of Terms of Service as Cyber-Regulation." *Internet Policy Review* 5(4)

[12] For a detailed analysis, albeit from 2015, of the 'Community Standards' and content moderation policies of Facebook and Twitter see e.g. P. Leerson, 'Cut Out By The Middleman: The Free Speech Implications Of Social Network Blocking and Banning In The EU' 6 (2015) JIPITEC 99.

[13] We also use this phrase to differentiate what we mean here from how the term 'self-regulation' is sometime defined in the literature, as a system in which 'the state provides the general objective, while the act of achieving that objective is performed by the private sector' (Benjamin Farrand, 'How do we understand online harms? The impact of conceptual divides on regulatory divergence between the Online Safety Act and Digital Services Act' (2024) *Journal of Media Law* 1, 8 and 3-6 https://doi.org/10.1080/17577632.2024.2357463 and the literature cited there, incl. Florian Saurwein, 'Regulatory Choice for Alternative Modes of Regulation: How Context Matters' (2011) 33 Law & Policy 334.

[14] Representative examples include Ronald Dworkin, "Foreword" in Ivan Hare and James Weinstein (Eds), EXTREME SPEECH AND DEMOCRACY (2010); James Weinstein, "Participatory democracy as the central value of American free speech doctrine," *Virginia Law Review* 97/3 (2011): 491-514; Eric Heinze, HATE SPEECH AND DEMOCRATIC CITIZENSHIP (2016).

As Evelyn Douek notes, this kind of voluntarily-sought expert input from civil society offered an attractive advance on simple self-regulation, commercially speaking: it allowed platforms to *show* that they were addressing the proliferation of dangerous falsehoods, without crossing the line of allowing government to dictate or illegitimately influence the boundaries of public debate.[15] We consider below (at §3.1) the significance of the recent retreat from such practices by some major platforms.

A further example of such enhanced self-regulatory practice is Facebook's *Redirect Initiative*, in which Facebook partners with terrorism diversion groups, in several countries, so that when users search Facebook for hate- and terror-related materials, they are directed towards content and organisations that provide anti-hate educational resources and support groups[16] – a notable instance of what has been analysed as platform-enabled counter-speech.[17]

Furthermore, civil society groups like the Electronic Frontier Foundation in the US, or the Open Rights Group in the UK, together with consumer pressure, has succeeded in persuading platforms to take measures to make their content moderation practices more transparent to public scrutiny, through various voluntary schemes of regular reporting and disclosures.[18] Facebook's Oversight Board is another self-regulatory initiative designed to provide a mechanism of review for major content- or speaker-removal decisions.[19]

However, even when making allowance for these enhanced forms of self-regulatory practice, serious legitimacy-related disadvantages with this model remain. First, strict abstention in this regulatory context may disempower government from upholding a minimal degree of material security and peaceful association. Depending purely on the individual policies of individual platforms, threats, disinformation, incitement, and stochastic terrorism, originating domestically or internationally, can all freely circulate in an unregulated discursive arena, in a way that at most poses risks to social and public

[15] E. Douek, "Content moderation as systems thinking," *Harvard LR* (2022) 136(2): 526-607, at 543-45.

[16] See https://counterspeech.fb.com/en/initiatives/redirect/ - the broken link may raise a question whether this initiative, too, is being abandoned, as part of Meta's tilt away from some previous content-moderation practices – below at 000-000.

[17] D. Citron and H Norton, 'Intermediaries and Hate Speech: Fostering Digital Citizenship for our Information Age' (2011) 91 *Boston Univ LR* 1435.

[18] Such policies require platforms to publish their moderation policies and regular twice-yearly transparency reports with aggregate quantitative data about their moderation activities (e.g., total number of posts removed). See generally, Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech.131 *Harv. L. Rev.* 1598-1670. Several bills have been brought before US Congress which would enshrine various transparency and reporting duties for platforms, but none of these bills have been passed; for discussion see *ibid.*

[19] Klonick, K. (2020). The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. 129 *Yale Law Journal* 2418.

order,[20] and at least, is sufficiently toxic and threatening to drive other speakers, especially women, off the major platforms[21] and even out of politics.[22] It is a dereliction of a government's duty to rely solely upon corporate self-regulation to manage these dangers.[23] It has allowed one extremely wealthy individual, Elon Musk, to personally reshape the guardrails of public discourse on Twitter/X, according to his own ideological outlook.[24]

Second, under CDA 230 and the First Amendment, while constitutional protection is very strong when online speech is suppressed by government action, it is non-existent when speech is suppressed or removed by the platform that hosts the speech.[25] Even where the state isn't actively suppressing its citizens' speech, it plausibly incurs a legitimacy deficit if its regulatory omissions risk allowing certain speakers being driven off platforms, by targeted threats and abuse. At minimum, a regime's democratic legitimacy can be undermined under this sort of system if there are no safeguards to try to prevent

[20] Examples in the US include most notoriously the January 6th insurrection at the Capitol for which Donald Trump was subsequently banned from Twitter and the 2024 summer riots in the UK, motivated by hostility to asylum-seekers fuelled by misinformation about a particularly shocking mass murder. In relation to the latter, Meta's oversight board voiced strong criticism of Meta's slow and flawed attempts to deal with multiple posts advocating violence on its platforms: https://news.sky.com/story/meta-too-slow-during-uk-riots-to-deal-with-violent-posts-online-says-independent-review-13354324

[21] Citron, n 17 above and, with J Penney, 'When Law Frees Us to Speak' (2019) 87(6) *Fordham LR* 2317.

[22] See e.g. 'Online vitriol could undo decades of political progress, warns Dutch deputy PM' *The Guardian* 3 November 2023 at https://www.theguardian.com/world/2023/nov/03/online-vitriol-could-undo-decades-political-progress-dutch-deputy-pm#:~:text=Six%20years%20after%20Sigrid%20Kaag,constant%20watch%20over%20her%20home

[23] In principle we could think of this as an additional criterion for assessing government legitimacy, e.g. something like maintaining law and order and security. But you might prefer to think of it as something even more basic. A government that can't maintain a minimal degree of material security – that's overwhelmed by outbreaks of lawless violence, or which has lost the capacity to police violent crime – is on verge of not being a government at all.

[24] For a recent analysis arguing that 'the organising principle' of Twitter/X is now 'its owner's whims', which has resulted in a collapse in moderation standards and a surge in disinformation on the site, see L. Kelley 'Elon Musk's Unrecognisable App' *The Atlantic* (1 November 2023) https://www.theatlantic.com/newsletters/archive/2023/11/elon-musk-year-twitter-x/675870/; see also Musk and X are epicentre of US election misinformation, experts say | Reuters.

[25] Dubbed 'a paradox' by leading free speech campaigner Nadine Strossen, "The Paradox of Free Speech in the Digital World: First Amendment Friendly Proposals for Promoting User Agency," 61 *Washburn LJ* 1-44 (2021). In response to concerns like these, Florida's SB 7072 Bill (2021) made it an offence for social media sites to deplatform political candidates and journalistic work, but this was found unconstitutional in *NetChoice LLC v Attorney General* of Florida 34 F.4th 1196 (11th Cir., 2022). By contrast, Texas's House Bill 20 (2021), restricting platforms and email clients from interfering with user's speech, was unexpectedly found constitutional in *NetChoice LLC v Paxton* (5th Cir., 2022). The decision of the Supreme Court (*Moody v. NetChoice, LLC and NetChoice, LLC v. Paxton*, 603 U.S. 707 (2024)) did not fully resolve the issue but strongly indicated the laws were both unconstitutional in a judgment that affirmed platform moderation choices to be protected by the First Amendment, thus following. Justice Kavanaugh's (*per curiam*) opinion, in *United States Telecom Association v. FCC*, 855 F.3d 381, 433-34 (DC Cir., 2017), that although "the real threat to free speech today comes from private entities such as Internet service providers, not from the Government," nevertheless, the government cannot "regulate the editorial decisions of Facebook and Google."

arbitrary exclusions of particular speakers or viewpoints from the major platforms,[26]such as the kind of robust mechanisms for speaker redress, we use as one of our five key criteria.[27]

This problem is exacerbated by the fact that content *demotion* (reducing speech's visibility via algorithmic tweaks, or adding 'friction' that makes sharing more difficult), rather than content removal, is often the expedient way for self-regulating platforms to manage speech that lies near the threshold of violating its own regulatory standards.[28] YouTube has said that such design decisions *a*re one or more orders of magnitude *more* important than content-removal.[29] But in a context in which private corporations are setting the ideological parameters of public discourse, systematic demotion may be worse than removal. If people's social media posts are being simply removed, they will very likely notice – *a fortiori* if they are kicked off the platform outright. But demotion can make disfavoured ideas *de facto* inconsequential by being rendered largely invisible in online public discourse, even while they remain *de jure* expressible, a process known as 'shadow-banning.' One journalist reporting on the so-called 'Twitter Files', said 'decisions to "actively limit the visibility of entire accounts or even trending topics" were made "in secret, without informing users."'[30] Demotion can turn curatorial prerogatives into a mechanism of covert ideological discourse-control.[31]

As we will note further below,[32] there is a powerful argument that, regardless of whether one approves or disapproves of the particular standards used for such different forms of content moderation, their impact is too important for their formulation to be left purely to the discretion of private corporations or single individuals. While the US First

[26] It is growing concern about this issue – albeit primarily from the political Right in US politics - that has led to the laws instanced *ibid.* Our point here is also partly influenced by Douek's view (n 15 above) that a *systems* approach is needed, moving away from the question of how a regulatory regime affects individual speakers, and towards how it conditions and structures the overall discursive environment.

[27] The problem has also led to calls for platforms to incorporate international human rights standards into their community guidelines (see e.g., David Kaye, The Global Struggle to Govern the Internet, https://globalreports.columbia.edu/books/speech-police/ and Applying International Human Rights Law for Use by Facebook, https://globalfreedomofexpression.columbia.edu/publications/applying-international-human-rights-law-for-use-by-facebook/; this would be a good thing (albeit perhaps looking rather less realistic, given recent events (below at 0000)) but still wholly the private choice of platforms.

[28] See e.g. Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, *8*(3) https://doi.org/10.1177/20563051221117552

[29] n 15 above, at 546.

[30] Reported by A Picchi 'Twitter Files: What they are and why they matter' https://www.cbsnews.com/news/twitter-files-matt-taibbi-bari-weiss-michael-shellenberger-elon-musk/

[31] Douek, n 15 above, at 545; on demotion see further Daphne Keller, "Who do you sue? State platform hybrid power over online speech," *Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1902* (January 29, 2019), www.lawfareblog.com/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech, at 17-19; for a defence of demotion as a potentially more proportionate way of dealing with some potentially harmful content, see Jeffrey Howard, Beatriz Kira and Louisa Bartolo, 'Remove or Reduce: Demotion, Content Moderation, and Human Rights' (2024) SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4891835

[32] Below at 000-000.

Amendment *requires* such abnegation on the basis that government suppression of speech on the basis of its content or viewpoint crosses a constitutional red line, the European perspective is the opposite: a purely private ordering of the online public sphere is *democratically* unacceptable; to put it more bluntly, self-regulation means democratic deficit. Interestingly this view is getting some purchase even in the US in the form of legislation at state level in Florida and Texas, which imposed 'keep up' requirements – seeking to prevent social media sites from deplatforming hosts or speakers on the basis of the content or message of their posts. While the Supreme Court has recent given support to the view that such interferences with the First Amendment rights of platforms to moderate as they see fit are unconstitutional,[33] they show a growing interest even in the US in the idea of subjecting content moderation to some form of democratic oversight.[34]

Third, an ironic consequence of the combination of the First Amendment and CDA 230, intended to guarantee the freedoms of platforms and speakers, is that it actually disables US law from intervening against platforms' decision to remove or delist controversial online content, even at the behest of sometimes highly authoritarian foreign governments.[35] Few things seem more emblematic of the dysfunctionality of the First Amendment in the digital age than this categorial prioritisation of the 'free speech rights' of global corporations like Meta and Google over that of the individuals whose speech they host. For a constitutional guarantee of free speech to open the door to online censorship at the behest of totalitarian China seems a particularly savage irony, vividly illuminating a structural weakness with *laissez-faire* regulatory strategies in a globalized world.

2.2 Punitive State Regulation

Model (E), Punitive State Regulation, is the operative model in EU law with respect to terrorism-related speech, under the *Regulation on Addressing the Dissemination of Terrorist Content Online* (TERREG).[36] In stark contrast to the US approach, these parts of EU law empower states to deal with at some harmful speech. The essence of this scheme is that a range of public authorities are empowered to notify for immediate removal *specific items of online content* that fall within the alarmingly broad and vague definition of 'terrorism-related content' in TERREG. Crucial for the classification of this as Model

[33] Above, n 25.

[34] The most recent, albeit narrowly-focused example was the 'Take it Down Act' passed by Congress in April 2025, criminalising publication of 'deep fake' sexual images or revenge porn and requiring social media sites to remove them on notice: edition.cnn.com/2025/04/28/politics/house-passes-bill-targeting-deepfake-revenge-pormelania-trump/index.html

[35] Keller (n 31 at 7ff) highlights two instances, *Zhang v. Baidu.Com*, 10 F. Supp. 3d 433 (S.D.N.Y. 2014), and *Sikhs for Justice v. Facebook*, 144 F. Supp. 3d 1088 (N.D. Cal. 2015); *Zhang v. Baidu* implicates a search engine rather than a social media platform; for the significance of this distinction, see H. Whitney and R. Simpson, "Search engines and free speech coverage" in S. Brison and K. Gelber, *Free Speech in the Digital Age* (Oxford: OUP, 2019) 33-51. See further A. Jaffe, "Digital shopping malls and state constitutions – a new font of free speech rights?" *Harvard Journal of Law & Technology* 33/1 (2019): 269, at 281-83.

[36] Regulation (EU) 2021/784 of 29 April 2021.This gave major additional powers to combat such content, supplementing those in the previous Directive 2017/541 on combating terrorism, OJ [2017] L 88/6.

(E) are two features: first, it relies on a broad definition that covers at least some content that was likely not already illegal in several Member States; second, the serious penalties the Regulation provides attach to failures to remove specific items of notified content, rather than (as with Model (D)) failures in overall process. It is the way the two factors work together that is so problematic. One could imagine a scheme like this being acceptable *if* it used a tightly-defined, narrow class of already unlawful, indisputably harmful, and easily identifiable content like child sexual abuse material (CSAM). Such content, if classed as speech at all, is scarcely a contribution to public discourse and a measure aimed at it would be unlikely to catch legitimate speech. But as will appear below, TERREG is nowhere near this kind of exceptional, possibly-acceptable hypothetical.

Again, the *prima facie* attractions of such Punitive State Regulation are evident. Consider the flipsides to our main objections to Self-Regulation. Whereas a refusal to regulate extreme speech online ignores the state's order- and security-related duties, Punitive State Regulation offers a way to fulfil those very duties. And whereas Self-Regulation lets companies (and thus at times, foreign powers) set the limits of public debate, Punitive State Regulation brings this under the control of the democratic polity. Moreover, unlike models (A)-(C) the scheme is EU public law; during its passage, it was publicly scrutinised, debated and amended in the European Parliament by MEPs elected by the citizens of all the Member States. So, it does answer to our criterion of democratic input. Thus, unlike the removal and demotion algorithms used by self-regulating platforms, or the opaque, informal decision-making involved in some hybrid schemes to be discussed below, the duties enshrined in Punitive State Regulation are a product of formal legislative processes, and depend on the policy decisions of identifiable governments agencies and actors.

Again, though, this approach also brings in a range of legitimacy-related problems. We can unpack these by considering several objections to the EU's TERREG laws, which are illustrative of the broader legitimacy-related worries around this whole approach.

First, TERREG's definition of terrorism content is an ill-defined jumble, which includes speech *advocating* and *glorifying* the commission of terror offences.[37] The worry here is simple. Do we trust government actors to distinguish – in a fair and principled way – unlawful incitement of terrorism, from lawful political radicalism ? As one of us has previously argued, the TERREG definition lumps together bomb-making instructions and direct incitement to violence with expressions of sympathy for groups using (or once using) violence, including defensive violence against oppressors. The latter is expression that is clearly on the spectrum of political speech. This reliance upon open-ended definitions risk imparting a dubious veneer of legality to the ideological persecution of dissenting factions, radical voices, or ethnic and religious minorities.[38] This is particularly

[37] TERREG Art 2(5)(a). There are exceptions for speech used in education, journalism, art, research, and awareness-raising. But these leave the central, problematic case untouched, in which radical political speech, used to try to rouse political action (not merely for art or awareness-raising etc.) is speciously misclassified as indirectly inciting, by glorifying, terrorism.

[38] Gavin Phillipson and Eliza Bechtold, "Glorifying censorship ? Anti-terror law, speech and online regulation" in Frederick Schauer and Adrienne Stone (Eds), OXFORD HANDBOOK ON FREEDOM OF SPEECH (2021): 518-41, at 521.

so given the growing prevalence in the EU of authoritarian governments, like those of Hungary. As one commentary has put it

> The fact that Member States with shameful human rights records and weakened rule of law will be able to delete online content throughout the EU clears the way for politically motivated censorship.[39]

Second, once authorities have issued removal orders to platforms, TERREG requires offending speech to be removed within *one hour,* across all platforms and members states, on pain of heavy fines (Article 3(3)). As the Danish civil liberties group Justicia has argued, it is totally unrealistic to expect 'thousands of complex speech complaints to be processed within hours, while simultaneously attaching proper weight to due process and freedom of expression.'[40] Any implementation of such a system that can operate at this speed and scale will be prone to suppressing speech in a relatively indiscriminate and context-insensitive fashion.[41] In other words TERREG scores particularly badly on our first criterion – abusability.

TERREG isn't totally insensitive to the concerns we are highlighting. It mandates the provision of an *effective remedy*, including a right to challenge removal orders before national courts. Removal orders must be notified to those whose material is being removed (Article 11), and be accompanied by justifying explanations (Article 3(4)(b)). Platforms must consider the complaints of those whose posts are removed expeditiously and if they got it wrong, restore the speech within two weeks (Article 10(2)). Moreover, member states must monitor their use of removal orders and make publicly available transparency reports detailing how much speech they remove, and instances of erroneous removal (Article 7). Such provisions do something to revive the legitimacy of such schemes in terms of our remaining criteria of transparency, accountability and redress.

We understand Douek's critique of regulators trying to 'have their cake and eat it too' by 'demand[ing] takedowns in ever-shorter timeframes without acknowledging the costs this will have for accuracy, sensitivity to context, or the practicality of individualized due process', which has particular force here.[42] We would stress, though, that such legal remedies are important at the very least for major moderation blunders - controversial decisions to remove high profile, or particularly valuable content. Keller highlights the case of an organization called Syrian Archive, which lost a cache of online videos that it was going to use in prosecuting human-rights abuses, after YouTube's content filters wrongly flagged them as terrorist content, due to their violent nature. In a state

[39] Chloé Berthélémy (EDRi), 'EU Terrorist Content Online Regulation Could Curtail Freedom of Expression across Europe' 3 February 2021 available at https://www.media-diversity.org/eu-terrorist-content-online-regulation-could-curtail-freedom-of-expression-across-europe/.

[40] Rushing to Judgment: Examining Government Mandated Content Moderation | Lawfare.

[41] Such a scheme also plainly contravenes the Manila Principles (above, n 9), in particular the requirement that platforms must not be required to restrict or remove content without a judicial order (Principle II), the UN Special Rapporteur Report (ibid), which imposes the same requirement, and the UN's Rabat Plan of Action on the obligation to outlaw hate speech under Article 20 ICPPR https://www.ohchr.org/en/freedom-of-expression.

[42] Douek, n 15 above, 566.

without legal remedies to regulate this action, all Syrian Archive could do was ask YouTube to reconsider.[43] When speech of the highest value both in political and evidential terms is erroneously removed like this, it seems completely inappropriate that the only possible recourse is a *request*, with zero legal force, made to a private commercial entity, that it may consider, at its sole discretion. The *legitimate* state will, at minimum, need to provide redress here, as a form of 'lateral' free speech protection. TERREG, to its credit, mandates this provision.

Nevertheless, the overall legitimacy-related worries remain, not just with TERREG, but with any Punitive State Regulation that empowers government to dictate content-removal *outcomes* to platforms based on broad and vague criteria. It is a commonplace in free speech theory that governments cannot, on pain of illegitimacy, invoke generic security concerns to justify the suppression of dissenting or otherwise politically disfavoured viewpoints. What is needed isn't just that governments refrain from this despotic sleight of hand, but that they have safeguards that make it harder for bad actors to abuse a potent regulatory apparatus in this way, and offer some kind of due process assurances to protect speakers in relation to the risks of arbitrary removal. Schemes like these sorely lack the first; they provide some protection for speakers in relation to the second, but only *ex post facto*. There is an in-built susceptibility, in this model, to the state's power to dictate content-removal *outcomes* being used in ways that run roughshod over the user's expressive rights and open the door wide to authoritarian abuse.

3. Classifying Midway Schemes

As our discussion so far indicates, we see both options (A) and (E) as having serious legitimacy-related drawbacks. We now turn to assess the legitimacy-related features of various Midway / Hybrid schemes, each of which, in principle, reflects a combination of Libertarian and Statist concerns. In this section we define and explain our three classes of Midway schemes, and then in §4 we offer a legitimacy-oriented evaluation of them.

3.1 Self-Regulation qualified by informal government influence.

We define Model (B), qualified Self-Regulation, as any scheme in which the self-regulatory content moderation practices of social media companies are *informally* influenced, pressured or even directed by government agencies.[44] Given platform power, it is wholly understandable that benevolent governments will try to at least 'advise' that moderation polices are exercised 'in ways that that promote the public good rather than just the companies' economic self-interest.'[45] The problems here are three-fold. First, governments

[43] Keller, n 31 above, at 7; Douek, ibid at 550.

[44] For a range of examples of government pressure on platforms and attempts by the Global Network Initiative to help platforms resist it, see n 12 above, paras 43-59.

[45] G Lakier Informal Government Coercion and The Problem of "Jawboning" - Lawfare (lawfareblog.com) (26 July 2021).

are seeking to obtain the *benefits* of influence over platform policy without accepting what should be the correlative *burden* of framing proposals for legislative deliberation, public debate and, ultimately, constitutional scrutiny in the courts. To that last point, it could be said the US case of *Missouri v Murthy* (formerly *Missouri v Murthy)* shows that, should the issue happen to come to public attention, there can in principle be constitutional challenge, albeit that the outcome of that particular case was inconclusive.[46] To which we would agree, but point out, second, that precisely because these are just informal practices, there are no built-in formal mechanisms for transparency and thus accountability; the press *might* find something out; but you might well also have covert influence operating behind closed doors that simply never come to light. Hence, third, precisely *because* such pressure may be covert and hence not subject to legal or political accountability, there are obvious dangers of partisan or authoritarian misuse.

All this has a further significant adverse implication. We said above that model (A), Self-Regulation, has been and is the *ostensible* operative model in the US. The problem, put simply, is that it can be hard, if not impossible to know when ostensible model A is actually model B. To give one example, we know that platforms increasingly collaborate with governments in maintaining databases of terrorist speech, using digital signatures (called *hashes*) to expedite the removal/deny the upload of known and reposted pro-terror content across all cooperating platforms.[47] One example of this is the Global Internet Forum to Counter Terrorism, which is an industry-led body that collaborates with state agencies and academic researchers to facilitate this kind of cooperative, cross-platform content removal.[48] Similar schemes are in place to facilitate and expedite the removal of child sexual abuse material (CSAM).[49] While we might applaud the aim, especially in relation to the latter, such measures are non-transparent ('Nobody apart from consortium members knows what is in the database or who added any piece of content'), lack redress ('there are no… independent mechanisms to audit or challenge inclusions'),[50] and greatly magnify the adverse impact on speakers – they may not realise their speech has been removed across *all* participating platforms, and 'will only be able to challenge the decision via each platform individually.'[51] Such schemes thus ramp up the

[46] A preliminary injunction issued by a US Federal Judge in Louisiana (4th July 2023), would have placed very substantial new limits on communications between government and platforms on First Amendment grounds (*Missouri v Biden,* Case 3:22-CV-01213). On appeal the Fifth Circuit Court of Appeals sharply limited the scope of prohibited communicative activities (No. 23-30445 (5th Cir. 2023)) but, as critical commentaries noted, failed to properly distinguish mere 'encouragement' from 'coercion'; see e.g. A Bhagwat https://knightcolumbia.org/blog/persuasion-or-coercion-the-fifth-circuits-muddled-view-of-missouri-v-biden. On appeal, the Supreme Court declined, on procedural grounds, to decide the case substantively: *Murthy v Missouri* 603 U.S. (2024).

[47] Jillian C. York and Ethan Zuckerman, "Moderating the public sphere" in Rikke Frank Jørgensen (Ed.), HUMAN RIGHTS IN THE AGE OF PLATFORMS (2019): 136-61, at 150.

[48] Keller (n 31 above, at 6) notes that, as at Jan 2018, 'twelve platforms were using a private database containing some 40,000 'hashes'.

[49] Douek, The Rise of Content Cartels, Knight First Amend. Inst (Feb. I1, 2020), https://knightcolumbia.org/content/the-rise-of-content-cartels

[50] Ibid at 24.

[51] Douek, n 15 above, at 543.

power and scope of content-removal decisions, while enabling government influence, but rendering the process both more draconian and more opaque to affected speakers.

While the above consist of voluntary cooperative schemes, there is also growing concern in the US at what is pejoratively called *jawboning*.[52] This compromises self-regulation, often in non-transparent ways, giving rise to the risk that the *appearance* of continuing platform control over content moderation practices masks the significant influence of state actors behind the scenes, advising, urging, or pressuring companies to moderate in line with government policy. Journalists publicising the 'Twitter files'[53] claim that they show a sustained volume of content-moderation demands from US government agencies for years and in particular in relation to those speakers the Biden and Trump administrations considered to be principally responsible for spreading vaccine misinformation during the COVID-19 pandemic.[54] The concern is that state power, exercised informally, via what has been called 'the invisible handshake',[55] can achieve a degree of government control over content moderation 'that would, if written into law, be struck down by courts' on First Amendment grounds.[56]

The obvious worry with this approach, in Keller's words, is that it elicits *anticipatory obedience* to government demands, enabling states to *de facto* decide when and how to restrict speech online, but without such actions being explicitly ratified in administrative or legislative acts that would then be subject to administrative and/or constitutional challenge.[57] First Amendment scholarship and literature seeks to distinguish between (i) merely informative or advisory input from government, resulting in a permissible kind of *persuasive* influence, and (ii) *coercive* pressure, in which companies face express or implied threats of (e.g. retaliatory regulation or adverse taxation) if they don't comply with government directives. The latter is generally seen as crossing the line into state action, hence constitutionally prohibited jawboning',[58] but scholars seem to disagree with each other about where the line should be drawn. For example, one writer notes how 'White House Press Secretary Jen Psaki announced that 12 people were producing 65 percent of the anti-vaccine misinformation on social media platforms and called on Facebook, in particular, to take "faster action against harmful posts."'[59] This sounds like category (i), but the writer cites it as an example of category (ii) jawboning. And, as noted

[52] For detailed analysis see Derek E. Bambauer, "Against Jawboning," (2015) 100(1) *Minnesota LR* 51-128.

[53] See above, n 30.

[54] See the allegations summarised https://nypost.com/2022/12/24/latest-batch-of-twitter-files-shows-cia-fbi-involved-in-content-moderation/ and discussed in detail in the judgments in n 46 above.

[55] M. Birnhack & N. Elkin-Koren, 'The Invisible Handshake: The Reemergence of the State in the Digital Environment', (2003 8(6) *Virginia Journal of Law and Technology* 8:6 (2003).

[56] Keller, n 31 at 5.

[57] Keller, ibid, at 3.

[58] Bambauer, n 52 above at 90, 92. And see above, n 46.

[59] Above, n 45.

above, the Supreme Court has recently declined to bring much needed clarity to the issue.[60]

These kinds of disagreement about exactly when (benevolent or acceptable) government *persuasion* shades into (illegitimate) government pressure illustrate the ways in which scholars and courts are conflicted about this issue. On the one hand, as noted above, governments may simply be trying to steer content moderation in ways that protect the public good. On the other, informal pressure risks allowing governments to do 'what the First Amendment is supposed to prevent them from doing' – using 'their economic and political power' to set the parameters of public discourse.[61]

This concern up to now has mainly focused on particular content-removal decisions (as with the Twitter files), but the 2024 election of Donald Trump to a second term saw a much more dramatic instance. Shortly after the election, Meta announced what have been described as 'sweeping changes to the content moderation policies' of its platforms[62] in a more libertarian direction – including the abandonment of fact-checking[63] – in a way that was expressly presented as alignment with the new administration's approach to free speech.[64]Meta's announcement even included a pledge to 'work with President Trump to push back on governments around the world', particularly those in Europe, said to be 'going after American companies and pushing to censor more'.[65] This *looks* like major government influence on platform policy, but exercised entirely behind the scenes.

Could it be argued that such a shift is at least indirectly democratic, as Mark Zuckerburg seemed to propose by claiming that Trump's 2024 election win 'feels like a cultural tipping point towards…prioritizing speech.'[66] The problem, again, lies in the obscure and informal nature of the influence at play here. There is no formal policy-making process, no public consultation, no role for democratic deliberation at all – simply an announcement by a corporation, which may or may not have been preceded by informal contacts between its senior figures and the incoming administration – or may have been a purely commercial calculation taken under the cloak of responsiveness to the new political *zeitgeist*. But assuming there *was* government influence here, the gravitational pull of Presidential power, yanking the polices of Meta in a plainly – and avowedly – Trumpian direction, is simultaneously extremely potent, but constitutionally invisible.

[60] Above, n 46

[61] Ibid.

[62] Meta's content moderation changes closely align with FIRE recommendations | The Foundation for Individual Rights and Expression

[63] 'We will end the current third party fact checking program in the United States and instead begin moving to a Community Notes program': Transcript: Mark Zuckerberg Announces Major Changes to Meta's Content Moderation Policies and Operations | TechPolicy.Press

[64] Ibid.

[65] Ibid.

[66] Ibid.

It is perhaps awareness of the particular risks of such informal power being exercised entirely non-transparently, 'in the shadows', that gave rise to some support for our next model, (C).

3.2 Non-Legislated Hybrid Schemes

Model (C) instantiates more formal schemes of jawboning. The difference is that, while they do not involve legislation, or the exercise of legal powers, these are formal schemes, arranged by public authorities and subject to at least some degree of transparency and hence more political (if not legal) accountability.

As an example of this option, consider the EU's Code of Practice on Disinformation,[67] and its Code of Conduct on Countering Illegal Hate Speech.[68] These are non-legal but formal, published agreements between the European Commission and a number of large social media platforms. These have become less significant in practical terms now we have the (Model (D)) EU Digital Services Act.[69] But this approach still deserves some consideration as an ostensibly more light-touch hybrid regulatory approach, eschewing the formal coercive powers of the state. We think however that this superficial appeal quickly dissipates on closer examination. Notably, the Hate Speech Code gives *de facto* powers to various government agencies in EU member states to notify the platforms of instances of illegal hate speech, which platforms then review – against criminal law, and their in-house community guidelines – rapidly removing the ones found to be unlawful, or non-guideline-compliant. We say 'rapidly', because in most cases – up to 70%+ of the time according to some Commission reports – notified content is removed within 24 hours. [70]

This system thus places quite strong levers to control extreme speech in the hands of national and supra-national authorities, even though they aren't backed-up by *de jure* legal duties or enforcement mechanisms. It does not, however, even attempt to give those whose posts are removed any mechanisms for due process or redress. While the Commission has given itself major influence over content-removal decisions, these schemes leave process and redress entirely in the hands of the platforms: hence anything speakers get in this respect, they would be offered anyway under self-regulation. Of course, a voluntary scheme *could* nudge the platforms towards providing better schemes

[67] The 2022 Code of Practice on Disinformation | Shaping Europe's digital future (europa.eu) This is broadly a set of commitments by major platforms to improve the content moderation of disinformation including demonetisation, empowering users, researchers and fact-checkers, prompt closure of false accounts etc. The Commitments of major platforms to it would seem to be in doubt given the developments at Meta and Twitter surveyed above.

[68] https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimi nation/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[69] The Disinformation Code has ceased to be a 'stand-alone' initiative by becoming integrated into the operation of the DSA: it will become 'a relevant benchmark for determining DSA compliance regarding disinformation risks' (https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation): see below at 000-000.

[70] See the Reports available at: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en#monitoringrounds.

for redress. Nevertheless, the fundamental objection to this model remains, similarly to Model (B), that the government is taking to itself quite effective controls over content moderation without placing those powers on a legislative basis and thus subjecting them to proper public, legislative and judicial scrutiny.[71] The scheme also scores poorly on democratic input (below, in §4.2).

However, in one respect, this model is preferable to Model (B); whereas government influence there may be hard to detect, or even deliberately concealed, the Codes are at least public documents, and the Commission publishes annual reports on their operation, which are thus open to public scrutiny.[72] The problem of jawboning is thus one that cuts across both Models (B) and (C), illustrating the normative complexity at play here. While the operation of arms-length governmental power is probably stronger in Model (C), its *relatively* more open status means it can score higher on transparency and lower on the risk of wholly *covert* government influence. On the other hand, the tools it gives government are probably much stronger. We evaluate this model further below.

3.3 Legislated Hybrid Schemes

This type of scheme, as we are conceiving of it, is exemplified by several European pieces of legislation, including the EU's *Digital Services Act* (DSA),[73] Germany's *Netzwerkdurchsetzungsgesetz* (NetzDG),[74] and the UK's *Online Safety Act* (2023).

These all involve *de jure* duties, backed by legal enforcement protocols and penalties. The crucial distinction between this model and Model (E), as exemplified in TERREG, is that whereas the latter empowers public authorities to specify specific content-removal *outcomes*, model (D) mandates broad content moderation *processes*. Under Model (D), states aren't primarily directing specific content-removal decisions. Instead, they are formalising the notification processes for flagging speech that's *already* unlawful, in a given jurisdiction, and then telling platforms what their processes must be for removing such content: how to make them transparent, consistent, and strongly amenable to complaints by, and remediation for, users.[75] As Uta Kohl says of NetzDG, the point of this

[71] Thereby breaches the Manila Principles. Principle VI b) (above, n 9) is that 'Governments must not use extra-judicial measures to restrict content [including] collateral pressures to…enforce so-called "voluntary" practices. While the Disinformation Code does not use the intrusive 'notice and takedown mechanism of the Hate Speech Code, it has still given the Commission considerable non-legislative influence over content moderation and hence, at least prior to its integration into the DSA, is vulnerable to our basic critique of Model C schemes.

[72] Above, n 70.

[73] EUR-Lex – 52020PC0825-EN-EUR-Lex (europa.eu) ('DSA').

[74] 1 Sept 2017, BGBl I S 3352).

[75] The UK's OSA exhaustively lists in Schedules 5-7 the illegal content in relation to which platforms have (per ss 9 & 10) a series of risk mitigation duties including removal; it also includes safety duties in relation to legal but harmful content aimed specifically at minors (ss 11-13) and 'user-empowerment' duties (ss 14-16). It prescribes particularly strong remedial schemes in relation to news publisher and journalistic content that has been removed: ss 18 & 19.

approach is to create “a public framework for private censorship, regulated self-regulation, in order to address systemic risks.”[76]

Consider the following four features of what is by far the most important of these schemes, in terms of cross-country reach, the DSA, which exemplify normatively significant features of this approach. First, the DSA retains the EU’s long-standing principle of conditional immunity. In essence, platforms are not obliged to perform comprehensive monitoring for illegal content.[77] They can only be held liable for a user’s speech if they have actual knowledge of the illegality of that content, by means of a formal *notice and take-down* under Article 9, or user notification under Article 16, and then only if they fail to remove this illegal speech ‘expeditiously’, upon receiving a precise notification of its illegality.[78]

Second, the DSA guarantees users’ due process rights around their speech, requiring (i) detailed explanations of content moderation decisions affecting users, which must (ii) include information about internal and external out-of-court avenues for seeking redress on any removal decisions[79] and (iii) ensure that user complaints are resolved under the supervision of qualified staff, not solely via automated processes (Article 20(6)).

Third, the DSA aims to radically boost platform transparency, requiring them to submit detailed annual reports on their content moderation policies and practices, which must cover both details of notice-and-take down orders received *and* details of the platforms own-initiative content moderation ‘that affects the availability, visibility and accessibility’ of posts[80] (thus also covering ‘demotion.’) And fourth, for the largest and most powerful social media and search companies, the DSA imposes additional risk-mitigation responsibilities, especially in relation to the risk of manipulation of their service, and its potential negative effects upon “the protection of public health, minors, civic discourse, or… electoral processes and public security.”[81] It is these broader duties that represent the DSA’s attempt at addressing more systemic and pervasive risks to democratic socie-

[76] Uta Kohl, “Platform regulation of hate speech – a transatlantic speech compromise?” *Journal of Media Law* 14/1 (2022): 25-49, at 37.

[77] DSA, Article 8.

[78] Article 6(1) and Recital 22. ‘Precise’ here meaning a notification that doesn’t necessitate further review or analysis. Article 9 sets out the content of such orders. Notification may also be by ‘trusted flaggers’ (Articles 16 & 22).

[79] Article 17 (and for Article 9 orders, Article 9(5)). Article 17 specifies that notifications to users of content-moderation decisions affecting them must include whether the action involved removal, demotion or demonetization… was in response to a notice submitted or based on voluntary own-initiative investigations, whether automation was used in the decision, a reference to the legal ground for identifying content as illegal or the relevant community guideline for a policy violation, and the redress mechanisms available to the user, including information about the possibility of out of-court dispute settlement provided for in Art. 21 as well as other available possibilities for redress (Article 20(5)), including through their domestic courts.

[80] Article 15 DSA.

[81] Article 34 and Recitals 81-90.

ties, including from disinformation and attempts at electoral interference. Much will depend on how actively the Commission exercises its powers here,[82] but this fourth feature is a particularly interesting one in encouraging platforms to think creatively about how to manage these more subtle and pervasive risks. Crucially, it means that 'Platforms' self-assessments and risk-mitigation efforts won't simply be taken on faith'.[83] Instead, they will be required 'to share their internal data with independent auditors, EU and Member State authorities, as well as researchers from academia and civil society', under Article 34. These bodies will provide independent scrutiny, to 'help identify systemic risks and hold platforms accountable for their obligation to rein them in.'[84] The platforms will then have to set out their plans to mitigate these risks, which will in turn be scrutinised by regulators (Article 35).[85]

How are such duties enforced? Under DSA Article 54, users can seek compensation from platforms for losses suffered as a result of non-compliance, and the DSA, NetzDG and OSA all allow for large financial penalties for systemic failure to comply with their mandates.[86] This obviously creates some potential for abuse or corruption, as in any other area where the state can levy fines upon media businesses. But we see these risks as being different from those that come with Punitive State Regulation, whose point, as we said in §2.2, is to massively extend and focus the state's speech-restrictive powers via the broad reach of platform moderation. The point of a Legislated Hybrid Scheme like the DSA, by contrast, is to get platform moderation policy to strike a reasonable balance between dual aims – restricting unlawful extreme speech, but also safeguarding and bolstering users' expressive interests, a point repeatedly stressed in the DSA.[87] Platforms are incentivised – partly through penalties – to enact good content-moderation, transparency and due process provisions: processes that reliably remove unlawful speech, but also ensure that lawful speech (which might well be removed under TERREG) *isn't* gratuitously penalised or disincentivized, and that there are effective remedies in respect of erroneous removal.

[82] The first such exercise was the announcement by the Commission of formal proceedings against X/Twitter on 18 December 2023 to investigate possible breaches of the DSA in relation to risk management, measures taken to combat information manipulation and transparency: see https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709

[83] 'A guide to the Digital Services Act, the EU's new law to rein in Big Tech' (AlgorithmWatch, 21 September 2022.

[84] Ibid.

[85] Martin Husovec stresses that these softer powers impose a clear 'red line' on the Commission: 'it cannot invent new binding content rules. That is, it cannot tell providers what *lawful* expressions they must prohibit or suppress on their services.' The Digital Service Act's Red Line: What the Commission Can and Cannot Do About Disinformation' (2024) SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4689926

[86] DSA., Articles 42 (Member States) and 59 (EU Commission). OSA, Chapter 6, Part 7 and Schedule 13.

[87] As stressed in the DSA in Article 14, and recitals 3, 22, 41, 47, 51-54, 63, 81, 86, 90, 153. For duties to safeguard expression and especially journalistic material under the UK's OSA, see above n 75.

4. The Case for Model (D)

4.1 Abusability, Transparency, Accountability and Redress.

All of models (B), (C), and (D), try to chart a 'midway' course, via cooperation between platforms and government, between pure self-regulation (A) on the one hand, and punitive state regulation (E) on the other. But such cooperation can create its own legitimacy problems by making it harder to assess exactly where and how states are responsible for the speech-restrictive policies and practices they are influencing. While all of the midway approaches are liable to criticism in this regard, we see Legislated Hybrid Schemes as better than the other two models, because they are more transparent and procedurally accountable, and less open to authoritarian abuse.

In terms of transparency, the issue with non-legislative forms of state action in this domain is that they allow, in Keller's words, government to *launder* censorious policy through platforms.[88] If state agencies informally direct public health advice on platforms, for example, this could allow states to suppress criticism of unpopular health policy, or bona fide dissent on e.g., lockdown efficacy, by classifying critical talking points as mis- or disinformation. Similarly, if state agencies 'advise' platforms to remove particular kinds of incendiary rhetoric, this may allow governments to suppress some hard-line dissident speech, but in a way that tends to elude political pushback, especially if the 'advice' is delivered covertly. Our point isn't just that it's bad for government to suppress lawful speech. The especially illegitimate thing we are highlighting is *extra-legal* suppression of speech, which fails to specify "clear requirements that could be put to judicial test."[89] Such schemes are a recipe for arms-length authoritarianism.

Moreover, because no legal powers are being exercised in Models (B) and (C), constitutional protections for speech aren't implicated, which thus inhibits the user's ability to legally contest the suppression of their speech. These approaches also tend to lack the kinds of remedies and safeguards – what Jack Balkin calls *curational due process* – that are essential to any legitimate regulatory policy pertaining to the exercise of core civil rights.[90] Whatever remedies there are to allow users to challenge speech-restrictive outcomes will operate at the discretion of platforms. Such schemes can thus function as semi-privatised systems of prior restraint, but without the legal requirements for notification and access to legal forms of redress that legislated restrictions would entail. They can thus allow for the *arbitrary removal* of presumptively lawful speech – acts of speech suppression which are, as Stefan Thiel characterizes them, (i) unilateral, (ii) based on the decision-maker's discretion, and (iii) not constrained by transparent rules.[91]

[88] Keller, n 15 above, at 3.

[89] Dinah Pokempner, "Regulating online speech: keeping humans, and human rights, at the core" in Brison and Kelber, n 35 above at 232.

[90] Jack Balkin, "Free speech is a triangle," COLUMBIA LAW REV (2018) 118)7) 2011-55, at 2040ff.

[91] S. Theil, 'Private censorship and Structural Dominance' (2022) 81(3) CAMBRIDGE LAW JO, 645, 666.

4.2 Democratic Input

Legislated Hybrid Schemes also provide for cooperative regulatory policy-making, between platforms and government, but in a way that mitigates these problems of transparency and due process. Our claims here dovetail with Utah Kohl's defence of NetzDG. Kohl argues that NetzDG adds a welcome degree of democratic input into policy decisions that set the parameters for public discourse, but which have been under the control of private companies.[92] NetzDG only prescribes the removal of already-unlawful speech, the contours of which have been set by the state's democratically legislated standards. Thus, on Kohl's view, NetzDG is merely augmenting the extant standards, imposed by global technology corporations, with "democratically legitimated community standards of the German polity."[93]

Kohl's argument anticipates and addresses a natural anxiety about the potential democratic *illegitimacy* of Legislated Hybrid Schemes. As we noted above, much recent philosophical work on free speech posits a tight link between free speech and democratic legitimacy.[94] On this view, even if a state confines itself to prescribing content moderation *procedures*, it is arguably still disrupting its people's upstream processes of debate and opinion-formation, and thereby undermining the democratic legitimacy of downstream electoral and legislative outcomes.[95] Kohl's point, however, is that given the immense, discourse-shaping power of contemporary social media, there is no organic, medium-independent process of opinion formation to be had. The best – the *most democratic* – thing we can do, under these conditions, is to ensure that social media's pervasive influence upon public discourse is organised by guidelines that have some meaningful democratic pedigree, instead of flowing simply from our corporate overlords' general commercial imperatives, or (*pace* Elon Musk)[96] their idiosyncratic personal agendas.

We agree with Kohl, as per our critical remarks about Self-Regulation above, (§2.1), but want to stress the differences between options (B), (C), and (D), in terms of how they link up with this reply to worries about democratic illegitimacy. All three of our Midway approaches involve cooperation between government and platforms. But it doesn't follow that they can *all* be defended as performing well in relation to our democratic input criterion. Different kinds of hybrid schemes may lend, or fail to lend, a democratic imprimatur to social media's structuring role in public discourse. The informal exercises of government influence on platforms that we see in Models (B) and (C) – coordinating, advising, urging, pressurizing – have a far more tenuous democratic pedigree than the types of influence exercised via Legislated Hybrid Schemes. Formal acts of legislation in a parliamentary democracy (including here the European Parliament) embody, however

[92] Luca Belli and Jamila Venturini, "Private ordering and the rise of terms of service as cyber-regulation," *Internet Policy Review* 5/4 (2016): 1-17.

[93] Kohl, n 76 above, at 36.

[94] Above n 14.

[95] The upstream/downstream terminology is borrowed from Dworkin, "Foreword" (ibid).

[96] And now perhaps Mark Zuckerberg: above, text to and following n 62.

imperfectly, the collective will of the polity. They articulate and implement that will in a way that is (however imperfectly) answerable to the polity in turn. Covert state action doesn't submit to public judgement, and this jeopardises its status as an expression of public will. Granted, model (C) avoids this particular problem by operating as a formal, published scheme. However, in being a product of the EU's 'executive' branch only, thus lacking meaningful democratic pedigree, (and offering nothing by way of due process and redress to speakers), it too must rank lower than (D), in relation to the criterion of democratic input.

4.3 The Aggrieved Speaker

How does all this look from the perspective of the aggrieved speaker – someone who wanted to express a controversial but lawful viewpoint, and who was obstructed because her chosen platform's content moderation policies removed or demoted her viewpoint? And how do considerations of government legitimacy relate to such conflicts between user and platform? What complaint, if any, might this person have against her government?[97]

In §2.1 we indicated our misgivings about the Libertarian/US constitutional response: that what private businesses like Facebook choose to moderate is none of the state's business. We think democratic governments can incur legitimacy deficits via regulatory *omissions* in this domain. Granted, one may deny this, e.g. if one thinks that, when it comes to speech-related harms, a government's legitimacy is unaffected by the dereliction of its harm-prevention duties. It would take us far afield to critique this view. But suffice it to say, we see it as un- (or under-) motivated. The right to free speech exists in order to support people's speech-related interests – including in inquiry, democratic debate, and the circulation of diverse perspectives.[98] In societies where private speech platforms play a pivotal role in facilitating public discourse, there is little to be said for a notion of speech rights that's based on a simple, two-way relation between speakers (rights-holders) and states (bearers of correlative duties).[99] Private speech platforms *do* have a right, presumptively, to host the speech they want to host. But democratic states

[97] A further possibility we do not consider here is that speakers may (at least in European polities) be able to enforce their free expression rights directly against private platforms: see Thiel, n 91 above.

[98] One might challenge such an interest-based view of expressive rights, arguing they are grounded in non-instrumental (e.g., dignity- or respect-based) duties of non-interference. But at this point in the dialectic that move would be self-defeating. Deontic justifications for speech rights can only justify rights held against interference by political authorities, or else they entail, absurdly, that a person's right to free speech is infringed any time her desired audience/platform declines to hear/host her expression. The idea that a speaker might be wronged, in being denied some platform, in a way that implicates state legitimacy (via regulatory absence or failure), requires an interests-based conception of free speech.

[99] Balkin, above, n 90 and "The future of free expression in a digital age," *Pepperdine Law Review* 36/2 (2009): 427-44; "Free speech in the algorithmic society: big data, private governance, and new school speech regulation," *UC Davis Law Review* 51/3 (2018); 1149-1210; "How to regulate (and not regulate) social media," *Journal of Free Speech Law* 1/1 (2021): 71-96.

have a duty to foster healthy public discourse,[100] and thus some justification for regulating platforms in a way that seeks to attain to this goal. While Facebook *is* a business, states have a prerogative, flowing from this duty, to responsibly regulate businesses like these that have significant *de facto* functions in shaping the democratic polis.

We think that our aggrieved speaker's most important complaint against her government – one that gives rise to a legitimacy deficit – would be about their government's failure to fulfil this secondary duty to regulate online platforms in a way that is more likely to protect healthy public discourse. Legislated Hybrid Schemes are more attractive in this way than non-legislative Midway approaches, because they put the public in a position to judge whether government is carrying out this duty. They also ensure that the aggrieved speaker is legally empowered to seek redress – if she believes her deplatforming was based on a misevaluation of her speech, or on procedural misapplication of a platform's own internal standards – rather than relying on the grace and favour of companies. And a further and final form of (legal) accountability *and* redress available under Model (D) lies in the fact that, if speakers and civil society groups believe that the state has taken regulatory powers that violate national or international guarantees of freedom of expression, they may challenge the relevant legislation itself before independent courts.[101]

5. Conclusion

One report on content moderation and freedom of expression argues that, "given the dangers of statutory regulation of content moderation, voluntary mechanisms between digital companies and public bodies represent a less intrusive, preferred approach."[102] The standard US First Amendment approach similarly insists that government's role must be limited to one of voluntary persuasion at most. This paper has argued that such critiques of government intervention are at best painting in unhelpfully broad brushstrokes. The dangers of statutory regulation around online speech aren't uniform,

[100] Such a duty finds expression in both UN and Council of Europe instruments. On the former, see UN Human Rights Committee, General Comment No. 34, (https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf) article 7, (explaining that the Article 19 ICCPR right to freedom of expression 'also requires States parties to ensure that persons are protected from any acts by private persons or entities that would impair the enjoyment of the freedoms of opinion and expression to the extent that these [ICCPR] rights are amenable to application between private persons or entities'.) On the latter, see Council of Europe, *Protecting the Right to Freedom of Expression under the European Convention on Human Rights,* section. 8.2 (on the positive obligation to protect the right) https://rm.coe.int/handbook-freedom-of-expression-eng/1680732814. For the growing academic discourse on this more active notion of free speech duties generally, see A. Kenyon and A Scott (eds) *Positive Free Speech: Rationales, Methods and Implications* (Hart: 2020).

[101] Given that the French Constitutional Council struck down most of the French AVIVA law Décision n° 2020-801 DC du 18 juin 2020 | Conseil constitutionnel (conseil-constitutionnel.fr)) challenges to schemes like TERREG and DSA (or their misuse) are a very real possibility. See Online hate in France - the 'Avia Law': the end of an intensive legislative saga (epra.org).

[102] 'The European Commission's Code of Conduct for Countering Illegal Hate Speech Online An analysis of freedom of expression implications' (2019), available at https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/EC_Code_of_Conduct_TWG_Bukovska_May_2019.pdf

across models (D) and (E). And it isn't true – either definitionally, or as a plausible empirical generalisation – that voluntary, cooperative regulation between online platforms and state agencies are less intrusive than other options. Government intervention under models (B) and (C) can be both highly intrusive (despite a lack of formal, enforceable duties) and less transparent than under models (D) and (E). And in principle, model (D) legislated schemes could be fairly light touch.

When it comes to broad normative distinctions between different models, we have argued that the key considerations are abusability, transparency, democratic input, accountability and due process, not simply the level of state involvement. The use of criminal law standards, repurposed from underpinning individual prosecutions to specifying categories of content to be removed, is a feature that can appear in all the models we have surveyed.[103] However, the application of these standards at such scale and extraordinary rapidity have to diagnose offending speech in generic, decontextualized ways, rather than via the explicit, jurisprudential application of doctrinal standards, and they inevitably automate the procedures of suppression, instead of imposing case-specific penalties on offenders. When states are working with companies on such processes, what matters, legitimacy-wise, is that their actions, and the outcomes they contribute to, can be observed, assessed, protested, and where appropriate, overturned either in the courts or by replacement legislation. The crux of our argument for Legislated Hybrid Schemes is that they safeguard these democratic capabilities.

Is this all meant to be a socio-legal prediction, then? Are we saying that Legislated Hybrid Schemes will do better than other kinds of Midway approaches in reducing the incidence of unjustly deplatformed speakers, or of states abusing their powers to regulate extreme speech in antidemocratic ways? No. Our argument is based on ideals of procedural legitimacy, in a broadly civic-republican spirit. State legitimacy isn't just about what government does, but what its constitutional architecture permits. Democratic legitimacy is contingent upon the establishment of legal and institutional safeguards that ensure the *modal robustness* of the citizenry's political liberties, by which we mean their utility not just in traditional scenarios against the state, but in a range of actual or possible scenarios including where the concern is with the decisions and policies of powerful private actors. The state has a serious legitimacy deficit if the only thing that prevents it from jawboning social media companies into quashing dissent is the upright character of its current office-holders. But it also invites authoritarian abuse if, as under Model (E), it gives itself powers to order immediate removal of specified content, especially under a scheme that defines removable content as alarmingly vaguely as TERREG, such that at least some clearly political speech may become vulnerable to such removal.

The state has good reasons to do what it can to promote robust public discourse on social media – to limit the harmful effects of extreme speech, while ensuring protections for the rights of speakers. We have argued that its *legitimacy* depends upon it doing this as

[103] It is clearly key to the schemes we surveyed in models (C) and (D); as noted above TERREG allows for removal of a broad class of material including much that is already criminal; platforms also removal material illegal in several or all states (e.g. CSAM) on their own initiative (Model (A) or on request from governments (Model (B).

transparently as possible, and in a way that embeds a culture of democratic and legal accountability with legally-guaranteed individual remedies.[104]

[104] Previous versions of this paper were workshopped at the Bonavero Institute of Human Rights, Oxford University and at the Information Society Project and the Law and Philosophy Speaker Series, both Yale Law School. Particular thanks are due to Tarun Khaitan for chairing the Bonavero event, Chinmayi Arun for arranging the ISP talk, Jack Balkin for chairing it and commenting extensively, and to Amin Ebrahimi Afrouzi, for organising and chairing the Law and Philosophy workshop and the participants at all the events; also, to Utah Kohl, Jake Rowbottom, Eliza Bechtold, Peter Coe, Ricki-Lee Gerbrandt, Daniel Joyce and Andrew Kenyon for reading and commenting on an earlier draft. We are also very grateful for the illuminating and helpful reports by the two anonymous reviewers for the Handbook.