PLATFORMED HATE SPEECH AGAINST WOMEN: BEYOND SELF-REGULATION

ANJALEE DE SILVA* AND CHRISTINE PARKER**

In this article, we argue that hate speech against women ('sex-based vilification') occurring on major social media platforms exists at the intersections of patriarchy and platform power and is thus platformed. Such speech is amplified as an aspect of platforms' instrumental power, and accommodated, and thus authorised, as an aspect of platforms' structural power. Platforms also seek to maintain control or influence over the conditions for their own regulation and governance through use of their discursive power. Relatedly, there is a privileging of self-regulatory action in current laws and law reform proposals for platform governance, which we argue means that platformed sex-based vilification is also auspiced by platforms. This auspicing, as an aspect of platforms' discursive power, represents an additional 'layer' of contempt for women, for which platforms currently are not, but should be, held accountable.

I INTRODUCTION

Communicative conduct expressing contempt for women is ubiquitous on major social media platforms.[1] It manifests in a range of formats and contexts,

* Anjalee de Silva is a Lecturer at Melbourne Law School, the University of Melbourne and an Associate Investigator at the University of Melbourne node of the ARC Centre of Excellence for Automated Decision-Making and Society ('ADM+S'). She is also a Fellow of the Women's Leadership Institute of Australia.

** Christine Parker is a Professor of Law at Melbourne Law School, the University of Melbourne and Chief Investigator and Lead at the University of Melbourne node of the ADM+S. This article has been written with the support of the ADM+S and the Women's Leadership Institute of Australia. It has benefited from the comments of various colleagues at the ADM+S, Melbourne Law School, and elsewhere, to whom we are grateful. Thank you in particular to colleagues attending the Civic Automated Decision-Making Seminar on 8 May 2023 at RMIT organised by Mark Andrejevic and the Law and Society Association of Australia and New Zealand Annual Conference 2023 in Sydney, at which drafts of this article were presented. Particular thanks as well to Jessica Tselepy, Rajesh Varghese, and Deylan Kilic-Aidani for their research assistance.

1 References to such conduct have been covered by *The Guardian*: 'The Web We Want', *The Guardian* (Web Series, 2016) <https://www.theguardian.com/technology/series/the-web-we-want>. See, eg, Emine Saner, 'Vile Online Abuse against Female MPs "Needs to be Challenged Now"', *The Guardian* (online, 18 June 2016) <https://www.theguardian.com/technology/2016/jun/18/vile-online-abuse-against-women-mps-needs-to-be-challenged-now>; Sandra Laville, 'Research Reveals Huge Scale of Social Media

including everything from casually sexist remarks to invective directed at female journalists and bloggers, speech characteristic of the 'Manosphere',[2] and 'revenge' pornography. The problem is especially apparent in the context of the cyber harassment of women.[3]

As part of campaigns of cyber harassment or otherwise, this kind of speech on platforms is often directed at women in positions of political leadership or with public profiles. In Australia, for example, female politicians across the political spectrum have spoken openly about their experiences in this regard. 'Ditch the witch' was infamously said of Julia Gillard while she was Labor Party Prime Minister.[4] The abuse she suffered on Twitter (now X) was still more misogynistic.[5] Mehreen Faruqi, a Greens Party Senator, has written candidly of the intersectional and especially vitriolic attacks she is subjected to on platforms as a Muslim woman of colour.[6] Sarah Hanson-Young, also a Greens Party Senator, has used humour to publicly confront some of the harassment she receives.[7]

Misogyny', *The Guardian* (online, 26 May 2016) <https://www.theguardian.com/technology/2016/may/25/yvette-cooper-leads-cross-party-campaign-against-online-abuse>. See below nn 27–35 and accompanying text for this article's definition of 'major social media platforms'. References to 'platforms' in this article are references to major social media platforms fitting this definition.

2 The 'Manosphere' is 'a broad term referring to a collection of online groups of men that share misogynistic and anti-feminist views towards women': 'Cracking the Code of the Manosphere', *Australian National University* (online, 7 March 2023) <https://www.anu.edu.au/news/all-news/cracking-the-code-of-the-manosphere>.

3 Danielle Keats Citron defines cyber harassment as 'involv[ing] the intentional infliction of substantial emotional distress accomplished by online speech that is persistent enough to amount to a "course of conduct" rather than an isolated incident', and defines cyber stalking as 'an online "course of conduct" that either causes a person to fear for his or her own safety or would cause a reasonable person to fear for his or her safety': Danielle Keats Citron, *Hate Crimes in Cyberspace* (Harvard University Press, 2014) 3 ('*Hate Crimes*'). For ease of reference, this article defines 'cyber harassment' as encompassing both cyber harassment and cyber stalking. Cyber harassment typically involves sustained and tactical campaigns of invective, image-based abuse, and other objectifying and derisory speech, and is often engaged in by 'cyber mobs' of more than one attacker. The relative invisibility of assailants online, as well as the multi- and cross-jurisdictional nature of cyber harassment, make it difficult to identify participants or measure the extent of any given mob. And though 'the totality of their actions inflicts devastating harm … the abuse cannot be pinned on a particular person': at 24.

4 James Massola and Political Correspondent, 'Julia Gillard on the Moment That Should Have Killed Tony Abbott's Career', *The Sydney Morning Herald* (online, 23 June 2015) <https://www.smh.com.au/politics/federal/julia-gillard-on-the-moment-that-should-have-killed-tony-abbotts-career-20150622-ghug63.html>.

5 Elle Hunt, Nick Evershed and Ri Liu, 'From Julia Gillard to Hillary Clinton: Online Abuse of Politicians around the World', *The Guardian* (online, 27 June 2016) <https://www.theguardian.com/technology/datablog/ng-interactive/2016/jun/27/from-julia-gillard-to-hillary-clinton-online-abuse-of-politicians-around-the-world>.

6 See, eg, Mehreen Faruqi, 'The Abuse and Hate I Get when I Speak Out Hurts: But Shutting Up Isn't an Option', *The Guardian* (online, 8 February 2019) <https://www.theguardian.com/commentisfree/2019/feb/08/the-abuse-and-hate-i-get-when-i-speak-out-hurts-but-shutting-up-isnt-an-option>.

7 See, eg, Sarah Hanson-Young, 'Pleasantries with Sarah Hanson-Young: Part 1' (YouTube, 19 March 2015) <https://www.youtube.com/watch?v=7HfBwVee6Ao>; Sarah Hanson-Young, 'Pleasantries with Sarah Hanson-Young: Part 2' (YouTube, 26 March 2015) <https://www.youtube.com/watch?v=3KKLN4J6gXc>; Sarah Hanson-Young, 'Senator Sarah Hanson-Young: "Insults Are a Daily Part of My Life"', *Mamamia* (online, 19 March 2015) <https://www.mamamia.com.au/sarah-hanson-young-talks-about-online-bullying>. Of course, YouTube is itself a platform, which highlights the double-edged nature of platforms' offerings.

International examples also abound, and women with public profiles may be particularly targeted on platforms when they speak openly about issues affecting women.[8] It is not only private actors responsible for such conduct; some governments and other state institutions have also employed cyber harassment against women on platforms to shut down dissent on feminist issues.[9]

The communicative conduct described often constitutes 'hate speech' against or the vilification of women, in the sense that it is about all women, even as it is directed at particular women. Noting the 'overwhelmingly impersonal, repetitive, stereotyped quality' of abuse that women receive online, Sady Doyle, for instance, argues that 'all of us are being called the same things, in the same tone'.[10] The communicative conduct described may be said to be directed at women *for being women*. We refer to such speech as 'sex-based vilification'.[11]

In this article, we argue that sex-based vilification occurring on platforms exists at the intersections of patriarchy and platform power and is platform*ed* in two main ways. First, communicative phenomena such as disinhibition, network and group polarisation effects, and (dis)information (or outrage) cascades are characteristic by-products of the affordances and infrastructures that form the core of platforms' value proposition and business models. These phenomena tend to *amplify* the prevalence and severity of some sex-based vilification on platforms, as an aspect of platforms' instrumental power, or ability to influence communication through control of social media.[12] Second, sex-based vilification may be *accommodated*, and thus *authorised*,

---

8 For example, Anita Sarkeesian, a Canadian American feminist blogger and gamer, was targeted with cyber harassment after starting a crowdfunding campaign to create a series of short films examining sexist stereotypes in video games: Emma Alice Jane, '"Back to the Kitchen, Cunt": Speaking the Unspeakable about Online Misogyny' (2014) 28(4) *Continuum* 558, 562. Caroline Criado-Perez was similarly besieged for heading up a successful campaign to have Jane Austen's image replace Charles Darwin's on the British £10 note. When Criado-Perez spoke out about the abuse, including during mainstream media interviews, the online campaign of invective against her escalated. Several high-profile women who pledged their support for Criado-Perez also received floods of abuse: at 563.

9 This recently occurred in Iran in response to protests surrounding the apparent arrest and murder by Iranian authorities of Mahsa Amini, allegedly for her failing to wear a hijab in public: Penny Wong, 'Joint Statement through the Global Partnership for Action on Gender-Based Online Harassment and Abuse on Standing with the Women and Girls of Iran' (Joint Media Statement, 9 December 2022) <https://www.foreignminister.gov.au/minister/penny-wong/media-release/joint-statement-through-global-partnership-action-gender-based-online-harassment-and-abuse-standing-women-and-girls-iran>. See, eg, Azin Mohajerin and Sussan Tahmasebi, 'Iranian Women's Rights Activists Face New Online Threats', *Global Voices Advox* (online, 25 August 2022) <https://advox.globalvoices.org/2022/08/25/iranian-womens-rights-activists-face-new-online-threats/>.

10 Sady Doyle, 'But How Do You Know It's Sexist? The #MenCallMeThings Round-Up', *Tiger Beatdown* (Blog Post, 10 November 2011) <http:www.tigerbeatdown.com/2011/11/10/but-how-do-you-know-its-sexist-the-mencallmethings-round-up/>.

11 We discuss this article's use of this term below at nn 19–23 and accompanying and surrounding text.

12 We discuss in detail what we mean when we refer to the instrumental, structural and discursive power of platforms below in Part III(A). Note that we are not primarily concerned with highly technical understandings of 'algorithmic amplification' or 'ranking' as discussed in literature: see, eg, Benjamin Laufer and Helen Nissenbaum, 'Algorithmic Displacement of Social Trust' (Essay, Knight First Amendment Institute at Columbia University, 29 November 2023) <https://knightcolumbia.org/content/algorithmic-displacement-of-social-trust>. Though this may be an aspect of amplification in the broader sense in which we use the term. Nor are we trying to suggest that the extent of the amplification (or accommodation or authorisation as discussed below in Part III(C)) of sex-based vilification on platforms

on platforms as a result of the relevant affordances and infrastructures. This is particularly significant because platforms now constitute, by design, something akin to a 'modern public square'.[13] Many users are compelled to participate in and on platforms for full social, economic and political inclusion and therefore have difficulty withdrawing from platforms. This is an aspect of platforms' structural power as significant channels for self-presentation and communication.

The platformed nature of sex-based vilification occurring on platforms thus impedes its regulation. Attempts to mitigate its harms to women are stymied by platforms' corporate and profit imperatives that underly its platforming and that conflict with, and often drive, platforms' responses to such speech. Significantly, platforms seek to maintain control or influence over the conditions for their own regulation and governance through use of their discursive power. Just as they make claims as to their economic, social and political value,[14] they also claim that (because they are so valuable) if they must be regulated, they themselves are uniquely placed to do the regulating. Platforms thus discursively constitute themselves as the (only or most) legitimate arbiters of the treatment of the speech they host.

Related to this is a privileging of self-regulatory action in current laws and law reform proposals for platform governance, which, as we argue, is both inadequate and inappropriate to mitigate the harms to women of platformed sex-based vilification. Self-regulation obfuscates the harms of platformed sex-based vilification by appearing to provide an internal governance solution that is not a solution and undermines potential for rendering platform power accountable for such speech. In doing so, it reinforces failures of existing laws to address such speech. It is also self-fulfilling, as the very existence of self-regulatory measures allows platforms to discursively construct themselves as *already regulated*, with a view to avoiding stricter, externally imposed oversight.

Through the use of their discursive power to privilege self-regulation, platforms are thus responsible for discursively 'rubberstamping' ineffective or anti-feminist content moderation processes and outcomes in ways that re-enact platformed sex-based vilification's harms to women, as well as platforms' complicity in those harms. We argue that this means that platformed sex-based vilification – being speech that is amplified, accommodated and authorised on platforms – is also *auspiced by* platforms, as an aspect of their discursive power. This auspicing

---

can be quantified with any accuracy. We are merely suggesting that speech constituting sex-based vilification appears to be amplified on platforms in the ways we describe in this article and that this is of normative concern in the context of its harms to women.

13 The Supreme Court of the United States recently went so far as to say that cyberspace is 'the most important place … for the exchange of views' and that social media sites constitute 'the modern public square': *Packingham v North Carolina*, 582 US 98, 99, 104 (2017).

14 See, eg, 'Meta Proudly Supports the People and Economy of the United States and around the World', *Meta* (Web Page) <https://research.facebook.com/economiccontribution/>; Kari Paul, 'Zuckerberg Defends Facebook as Bastion of "Free Expression" in Speech', *The Guardian* (online, 18 October 2019) <https://www.theguardian.com/technology/2019/oct/17/mark-zuckerberg-facebook-free-expression-speech>; Andrew Marantz, 'Facebook and the "Free Speech" Excuse', *The New Yorker* (online, 31 October 2019) <https://www.newyorker.com/news/daily-comment/facebook-and-the-free-speech-excuse>.

represents an additional manifestation or 'layer' of contempt for women, for which platforms currently are not but should be critiqued and held accountable.

Throughout this article, we employ a theoretical framework combining a multifaceted analysis of platform power[15] with a critical feminist understanding of sex-based vilification as discriminatory conduct constituting and causing the systemic subordination and silencing of women.[16] We refer to a range of regulatory and governance responses, including domestic and transnational anti-vilification laws, state endorsed voluntary and quasi self-regulatory content moderation standards, and self-regulatory and organisation specific content moderation policies. Many of these are complex and create multiple regulatory schemes with respect to different types of online content. For clarity, we refer only to those aspects of these measures that are most relevant to the regulation of platformed sex-based vilification.[17] Our focus is on the Australian, United Kingdom ('UK') and European contexts. These are jurisdictions that take broadly similar approaches to the regulation of vilifying speech through law and in which recent and significant legislative developments have occurred with respect to platform regulation.[18]

We also use 'sex' and 'sex-based vilification' in favour of 'gender' and 'gender(ed) vilification' deliberately. Sex as referred to throughout this article includes actual and perceived sex. Sex-based vilification includes speech directed at both cis and transwomen on the basis of their actual or perceived female sex. Much communicative conduct that might be characterised as speech vilifying women is *explicitly* sex-based.[19] Sex-based vilification is also distinct from vilification on the basis of gender identity, as is prohibited in some Australian states and territories.[20]

---

15 Doris Fuchs, 'Theorizing the Power of Global Companies' in John Mikler (ed), *The Handbook of Global Companies* (Wiley-Blackwell, 2013) 77 <https://doi.org/10.1002/9781118326152.ch5> ('Theorizing Power'). We discuss in detail what we mean when we refer to the instrumental, structural and discursive power of platforms below in Part III(A).

16 Anjalee de Silva, 'Addressing the Vilification of Women: A Functional Theory of Harm and Implications for Law' (2020) 43(3) *Melbourne University Law Review* 987. We discuss this functional theory of the harms of sex-based vilification in detail below in Part II.

17 Relatedly, as child protection is a specific issue outside the scope of this article, we only discuss these measures as they relate to adult platform users.

18 These are, primarily: *Online Safety Act 2021* (Cth) ('*AU Safety Act*'); *Online Safety Act 2023* (UK) ('*UK Safety Act*'); *Regulation* (*EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act)* [2022] OJ L 277/1 ('*Digital Services Act*'). These are discussed below in Part IV(A)(2).

19 For example, threats of sexual violence employed as part of the cyber harassment of women often refer in graphic detail to their (imagined) female bodies. Related to this is that prevailing norms in patriarchal societies mean that many women – both cis and trans – are regularly presumed to be cis unless they are known or perceived to be trans. Louise Richardson-Self notes, for example, that 'under patriarchy, being cisgender is taken for granted': Louise Richardson-Self, *Hate Speech Against Women Online: Concepts and Countermeasures* (Rowman & Littlefield, 2021) 35.

20 In the Australian context, see, eg, *Criminal Code 2002* (ACT) s 750 ('*ACT Criminal Code*'); *Discrimination Act 1991* (ACT) s 67A ('*ACT Discrimination Act*'); *Crimes Act 1900* (NSW) s 93Z ('*NSW Crimes Act*'); *Anti-Discrimination Act 1991* (Qld) s 124A ('*Qld Anti-Discrimination Act*'). The recently enacted *UK Safety Act*, which we discuss below in Part III, refers to 'sex' and 'gender reassignment' but not 'gender': see, eg, *UK Safety Act* (n 18) ss 16, 62. The *AU Safety Act* (n 18) and the *Digital Services Act* (n 18), which we also discuss in Part III, are not framed in terms of protected characteristics at all. The Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill

This category of vilification has traditionally been addressed to vilifying speech directed at and about transgender and intersex persons for being transgender or intersex and typically excludes vilification directed at and about women, including transwomen, for being *women*.[21]

The term 'sex-based vilification' thus performs the dual functions of providing both conceptual and terminological clarity. Importantly, it also does the work of elucidating that transwomen can experience vilifying speech intersectionally, both on the basis of their gender identity and perceived female sex, depending on the context. Sex-based vilification occurring on platforms is also experienced by women intersectionally on the bases of other attributes such as race, religion, sexuality, disability, class and so on.[22] Women of colour and lesbian women, for instance, are frequently targeted differently and particularly virulently, in terms of both the nature and quantity of sex-based vilification to which they are subjected.[23]

We also use the term 'vilification' throughout this article deliberately, in contrast to much of the existing 'hate speech' literature. This is because the latter term tends to shift focus from the functions of discriminatory speech to its expressive qualities and in this way can be misleading. In fact, it is not the 'hate' in hate speech that is of relevance. As we extrapolate below, what is of concern is what such speech *does*.

Finally, we define 'social media platforms' along three dimensions. First, they function as internet-based *communication channels* 'that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others'.[24] Second, their principal means of value creation is in platforming communication and interaction in a 'system'[25] comprising of the platform owner and an array of autonomous complementors and consumers. That is, as flagged above, the platform owner enables social, economic, and increasingly political interaction through the provision and maintenance of a technical infrastructure and a set of governance

2023 (Cth) ('Misinformation and Disinformation Bill'), which we also discuss below in Part III, refers to 'gender' only: see, eg, at s 2 (definition of 'harm').

21 In the Australian context, 'gender identity' seems primarily to be intended to replace more outdated language referring to trans people and excludes cis women by implication. With respect to the *NSW Crimes Act* (n 20) s 93Z: see, eg, New South Wales, *Parliamentary Debates*, Legislative Assembly, 5 June 2018, 42 (Mark Speakman, Attorney-General).

22 Intersectionality is a theoretical framework widely attributed to Kimberlé Crenshaw. For an account of harmful speech at the intersections of sex and race: Kimberlé Williams Crenshaw, 'Beyond Racism and Misogyny: Black Feminism and 2 Live Crew' in Mari J Matsuda et al (eds), *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (Westview Press, 1993) 111.

23 See, eg, Citron, *Hate Crimes* (n 3) 13–16. See also Danielle Keats Citron, 'Law's Expressive Value in Combating Cyber Gender Harassment' (2009) 108(3) *Michigan Law Review* 373, 380.

24 Caleb T Carr and Rebecca A Hayes, 'Social Media: Defining, Developing, and Divining' (2015) 23(1) *Atlantic Journal of Communication* 46, 50 <https://doi.org/10.1080/15456870.2015.972282>.

25 Jennifer Cobbe, Michael Veale and Jatinder Singh, 'Understanding Accountability in Algorithmic Supply Chains' (Conference Paper, ACM Conference on Fairness, Accountability, and Transparency, 12–15 June 2023) <https://doi.org/10.31235/osf.io/p4sey>.

mechanisms that are designed to provide value for users but also, and primarily, serve the commercial interests of the platform owner.[26]

Third, because we are investigating the power of social media platforms in the digital speech economy, we focus on *major* social media platforms. We adopt a pragmatic definition of major social media platforms in line with the European Union ('EU') *Digital Services Act*,[27] which places extra responsibilities for due diligence in relation to harmful content on 'very large' online platforms ('VLOPs').[28] These are defined as those platforms and search engines used by more than 10% of consumers in the EU.[29] The first set of social media platforms captured by the *Digital Services Act* as VLOPs were Facebook, Instagram, LinkedIn, Pinterest, Snapchat, TikTok, Twitter/X and YouTube.[30] Many smaller platforms follow similar affordances and protocols to the major social media platforms, but there are some that intentionally differentiate themselves and seek to provide alternate spaces for discourse.[31] Overall, our three-pronged definition emphasises major social media platforms' situation as key spaces of discourse and engagement within public life with saliences that make them a significant locus of critical inquiry into legal accountability for sex-based vilification.

Within the above definition, we focus our analysis on Facebook. Facebook is emblematic of major social media platforms. With billions of active users accessing the platform in countries around the world in over 100 languages,[32] it is one of the most prominent platforms in the world and is owned by Meta,

---

26 Andreas Hein et al, 'Digital Platform Ecosystems' (2020) 30(1) *Electronic Markets* 87, 90 <https://doi.org/10.1007/s12525-019-00377-4>.

27 *Digital Services Act* (n 18).

28 The *Digital Services Act* (n 18) applies to all platforms (meaning 'platforms' in its common usage) since early 2024 but applied to VLOPs from August 2023: 'The Digital Services Act', *European Commission* (Web Page, 2024) <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en>.

29 *Digital Services Act* (n 18) art 33.

30 European Commission, 'Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines' (Press Release, 25 April 2023) <https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413>. The retail platforms Alibaba AliExpress, Amazon Store, Apple AppStore, Booking.com, Google Play, Google Shopping, and Zalando are also captured, as are Google Maps and Wikipedia. Facebook in both Australia and the UK would be captured using this approach. See Jemma Healy, '2024 Social Media Statistics for Australia', *Meltwater* (Blog Post, 29 April 2024) <https://www.meltwater.com/en/blog/social-media-statistics-australia>; Jess Smith, 'UK Social Media Statistics [Updated 2023]', *Meltwater* (Blog Post, 22 February 2023) <https://www.meltwater.com/en/blog/uk-social-media-statistics>. The Commission has since designated a second set of VLOPs under the *Digital Services Act*, however, these all relate to adult content sites: see European Commission, 'Commission Designates Second Set of Very Large Online Platforms under the Digital Services Act' (Press Release, 20 December 2023) <https://digital-strategy.ec.europa.eu/en/news/commission-designates-second-set-very-large-online-platforms-under-digital-services-act>.

31 The Signal messaging platform, for example, seeks to differentiate itself on privacy and security markers: *Signal* (Website) <https://signal.org/>.

32 Evelyn Douek, 'Facebook's "Oversight Board": Move Fast with Stable Infrastructure and Humility' (2019) 21(1) *North Carolina Journal of Law and Technology* 1, 4. See also Tom Simonite, 'Facebook Is Everywhere; Its Moderation Is Nowhere Close', *Wired* (online, 25 October 2021) <https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/>.

arguably the most powerful social media company in the world.[33] Through its establishment of the Facebook Oversight Board ('FOB'), which Evelyn Douek, for instance, describes as 'one of the most ambitious constitutional projects of the modern era',[34] Facebook is also uniquely placed to serve as an example of the role of platforms' discursive power.[35] It must be noted at the outset that this article was substantively finalised prior to Meta's decision in early 2025 to significantly depart from Facebook's pre-existing content moderation practices and move instead to a 'Community Notes' model similar to that employed on X.[36] This move was accompanied by a relaxing of Facebook's content moderation policies, including its 'hateful conduct' policy.[37] We anticipate that these changes will result in the further amplification, accommodation and authorisation of sex-based vilification on Facebook, in the ways discussed in this article, and the resultant exacerbation of the relevant subordination and silencing that cause harms to women.[38] The changes also further demonstrate that platforms privilege and frame self-regulation in ways

33 Meta Platforms owns four of the largest social media platforms, being Facebook, WhatsApp, Facebook Messenger, and Instagram: 'Most Popular Social Networks Worldwide as of February 2025, Ranked by Number of Monthly Active Users', *Statista* (Web Page, February 2025) <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>.

34 Douek (n 32) 2. See also Josh Cowls et al, 'Constitutional Metaphors: Facebook's "Supreme Court" and the Legitimation of Platform Governance' (2024) 26(5) *New Media and Society* 2448 <https://doi.org/10.1177/14614448221085559>.

35 That is, the establishment of the FOB is itself an exercise in discursive power legitimating platform self-regulation. See also below nn 238–41 and accompanying and surrounding text. It is unclear what form the work of the FOB will take in light of recent significant changes to Facebook's content moderation practices. See below nn 36–9 and accompanying text.

36 Joel Kaplan, 'More Speech and Fewer Mistakes', *Meta* (Web Page, 7 January 2025) <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>.

37 See, eg, Robert Booth, 'Revisions of "Hateful Conduct": What Users Can Now Say on Meta Platforms', *The Guardian* (online, 9 January 2025) <https://www.theguardian.com/technology/2025/jan/08/permitted-hateful-conduct-what-users-can-now-say-on-meta-platforms>; Freya Jetson, 'LGBTQ+ Advocates Alarmed by Meta's Hateful Conduct Policy Changes', *ABC News* (online, 10 January 2025) <https://www.abc.net.au/news/2025-01-10/meta-hateful-conduct-policy-changes-alarm-lgbtq-advocates/104800042>; Clare Duffy, 'Calling Women "Household Objects" Now Permitted on Facebook after Meta Updated Its Guidelines', *CNN Business* (online, 8 January 2025) <https://edition.cnn.com/2025/01/07/tech/meta-hateful-conduct-policy-update-fact-check/index.html>. Facebook's previous 'hate speech' policy (version as at 29 February 2024) is discussed below in Parts IV(C) and V of this article.

38 See 'Meta's Misinformation Shift: A Decision to Ditch the Experts with Silvia Montana-Nino', *Automated Societies* (ADM+S Centre, 4 February 2025) <https://podcasts.apple.com/fj/podcast/metas-misinformation-shift-a-decision-to-ditch/id1571619455?i=1000688702675> ('Meta's Misinformation Shift'). See also Ned Watt, Michelle Riedlinger and Silvia Montaña-Niño, 'Meta is Abandoning Fact Checking: This Doesn't Bode Well for the Fight against Misinformation', *The Conversation* (online, 8 January 2025) <https://theconversation.com/meta-is-abandoning-fact-checking-this-doesnt-bode-well-for-the-fight-against-misinformation-246878>. Research from the Center for Countering Digital Hate has previously highlighted the ineffectiveness of X's community notes model, including in relation to hateful expression, which can go hand in hand with misinformation in more general terms: 'X Ran Ads on Five Accounts Pushing Lies and Hate during UK Riots', *Center for Countering Digital Hate* (Web Page, 19 August 2024) <https://counterhate.com/research/x-ran-ads-on-five-accounts-pushing-lies-and-hate-during-uk-riots>.

that consolidate their power,[39] in line with our arguments regarding the auspicing of sex-based vilification on platforms.

In Part II, drawing on a critical feminist framework, we begin by outlining what we mean when we say that sex-based vilification is discriminatory conduct that constitutes and causes the systemic subordination and silencing of women. We argue that such speech harms women individually and as a group and that the harms are partly democratic harms.

In Part III, we introduce Doris Fuchs' three-fold typology of the instrumental, structural and discursive power of corporations (Part III(A)), which we refer to throughout the remainder of this article. We apply this typology in this section to show how sex-based vilification occurring on platforms may be said to be platformed in the first two ways described above. First, sex-based vilification is speech that is amplified on platforms as an aspect of platforms' instrumental power (Part III(B)). Second, it is speech that is accommodated and thus authorised on platforms as an aspect of platforms' structural or infrastructural power (Part III(C)). (The third face of platform power, discursive power, is discussed in Part V.)

In Part IV, we consider existing legal responses to platformed sex-based vilification, with a focus on platform liability (Part IV(A)). We show that there is a 'gap' in the law, which leaves women unprotected from the systemic subordination and silencing harms of such speech (Part IV(B)). Accordingly, the main regulatory responses that construct platform responsibility for sex-based vilification privilege self-regulation, in the form of organisation specific content moderation policies and practices. We show with reference to Facebook's internal 'Community Standards' on hateful conduct[40] that these measures are neither adequate nor appropriate to address the systemic subordination and silencing harms to women of platformed sex-based vilification (Part IV(C)).

In Part V, we argue that the privileging of self-regulatory measures of the kind discussed in Part IV(C) is related to platforms' exercise of their discursive power to protect themselves from external answerability. With reference in particular to the FOB, we show that such measures obfuscate the harms of platformed sex-based vilification and reinforce failures of existing laws to address such speech. We argue that platforms' use of their discursive power to seek to privilege self-regulation, as well as the discursive power platforms wield *through* self-regulation, thus re-enact platformed sex-based vilification's harms to women, as well as platforms'

---

39 For example, these most recent changes to Facebook's content moderation policies and practices appear to be a politically and ideologically motivated attempt by Meta to align itself with the Trump administration: see, eg, 'A New Facebook for the Era of President Trump: Podcast', *Today in Focus* (The Guardian, 14 January 2025) <https://www.theguardian.com/news/audio/2025/jan/14/a-new-facebook-for-the-era-of-president-trump-podcast>. See also Justin Hendrix, 'Transcript: Mark Zuckerberg Announces Major Changes to Meta's Content Moderation Policies and Operations', *Tech Policy.Press* (Web Page, 8 January 2025) <https://www.techpolicy.press/transcript-mark-zuckerberg-announces-major-changes-to-metas-content-moderation-policies-and-operations/>.

40 'Hateful Conduct', *Meta* (Web Page, 7 January 2025) <https://transparency.fb.com/en-gb/policies/community-standards/hateful-conduct/> ('Hateful Conduct Policy'). See also above nn 36–8 and accompanying text.

complicity in those harms. We argue that this represents an auspicing by platforms of such speech, for which platforms ought to be held accountable.

II HARMS OF SEX-BASED VILIFICATION

Vilifying speech harms systemically, through its relationship to systemic oppression and its ability to establish and reinforce structurally unjust social norms.[41] In this vein, we adopt a conceptualisation of sex-based vilification as discriminatory conduct that constitutes and causes the subordination and silencing of women.[42]

Sex-based vilification's subordination and silencing harms are systemic for two reasons. First, as introduced above, these harms accrue to women on the basis of their actual or perceived female sex, which is an axis of structural discrimination and disadvantage in patriarchal societies.[43] Second, speakers who engage in sex-based vilification have what may be described as 'covert' authority[44] in patriarchal societies. Speakers play by the rules of patriarchy when they engage in speech acts of sex-based vilification and are able to (re)enact its 'permissibility facts'[45] or the structurally unjust social norms of which it is comprised. For example, some vilifying speech directed at or about women treats women as sexual objects or as otherwise inferior on the basis of their sex. In so treating, and because its speakers have authority in patriarchal societies, such speech ranks women accordingly and legitimates or normalises that same treatment. In these and other ways, women are systemically constituted and reconstituted by sex-based vilification as subordinate.[46] They are also constituted and reconstituted as silenced, meaning as having no business speaking or nothing of consequence to say. Sex-based vilification silences women by preventing them from speaking, certainly, as discussed below. But it

41 Anjalee de Silva and Robert Mark Simpson, 'Law as Counterspeech' (2023) 26(4) *Ethical Theory and Moral Practice* 493, 497 <https://doi.org/10.1007/s10677-022-10335-3>.

42 See above n 16. This framing draws on the work of feminist and critical race scholars on oppressive speech, as well as speech act theory and related work on linguistic pragmatics, particularly of Rae Langton and Mary Kate McGowan, as referenced throughout. As with any theoretical contribution, it is an account that is contestable. However, its essential underpinning, that vilifying speech impacts on social norms in ways that are harmful to its targets, has relatively broad-church support in the hate speech literature. For example, outside the critical tradition, Jeremy Waldron claims that some pornography constituting sex-based vilification according to our definition has a pervasive 'pedagogical function': Jeremy Waldron, *The Harm in Hate Speech* (Harvard University Press, 2012) 90–2. See also Alexander Tsesis, *Destructive Messages: How Hate Speech Paves the Way for Harmful Social Movements* (New York University Press, 2002).

43 Male sex, on the other hand, is not an axis of structural discrimination and disadvantage in patriarchal societies. For this reason, contemptuous speech directed at and about men and boys on the basis of their male sex does not and cannot systemically harm them in the ways that sex-based vilification harms women in such societies. Vilification as experienced by gender diverse and non-binary people is outside the scope of this article.

44 Mary Kate McGowan, 'Oppressive Speech' (2009) 87(3) *Australasian Journal of Philosophy* 389, 395–7 <https://doi.org/10.1080/00048400802370334>.

45 See de Silva (n 16) 1021. See also ibid.

46 See de Silva (n 16) 1020–3.

also silences women by marginalising and devaluing their speech and in building structural constraints on their speech. The result is that even where women can and do speak, what they say is often unable to have its intended force.[47]

Causal harms of systemic subordination and silencing may follow sex-based vilification's constitutive harms. Permissibility facts enacted through speech acts of sex-based vilification may be accommodated as 'correct play'.[48] Conversely, they may be challenged or 'undone', for example by counter speakers.[49] Where permissibility facts are accommodated, they alter normative facts about what is permissible and possible in patriarchal oppression, meaning what is permissible and possible in relation to the treatment of women in patriarchal societies per se.[50] As hearers' beliefs, desires, and other emotions tend to accommodate to these shifts,[51] hearers' attitudes may also evolve to accord with permissibility facts enacted through sex-based vilification. Or their preferences may be triggered or conditioned by those permissibility facts.[52] These shifts in turn render it more likely that hearers will act on those permissibility facts.[53] Thus, speech acts of sex-based vilification may cause, in addition to constitute, the systemic subordination and silencing of women in patriarchal societies and may be seen to contribute to both discrimination and violence against women.[54] Causal harms may also manifest more directly. Sex-based vilification typically causes women to feel threatened and humiliated and to adapt their own behaviours accordingly, such as by policing their identities, speech and movements or by leaving online and offline spaces and disengaging from public life.[55]

This is not to say that sex-based vilification alone or even primarily constitutes women's oppression in patriarchal societies. Material factors such as male violence against women and disparities in employment, pay, property ownership and caring

---

47 Ibid.

48 David Lewis, 'Scorekeeping in a Language Game' (1979) 8(1) *Journal of Philosophical Logic* 339, 342–4 <https://doi.org/10.1007/BF00258436>.

49 Rae Langton, 'Beyond Belief: Pragmatics in Hate Speech and Pornography' in Ishani Maitra and Mary Kate McGowan (eds), *Speech and Harm: Controversies over Free Speech* (Oxford University Press, 2012) 72, 83–4 <https://doi.org/10.1093/acprof:oso/9780199236282.003.0004> ('Beyond Belief'); Rae Langton, 'How to Undo Things with Words' (John Locke Lecture, University of Oxford, 3 June 2015) <https://media.philosophy.ox.ac.uk/assets/mp3_file/0010/37387/Lecture_6.mp3> ('How to Undo'); Rae Langton, 'Blocking as Counter-Speech' in Daniel Fogal, Daniel W Harris and Matt Moss (eds), *New Work on Speech Acts* (Oxford University Press, 2018) 144 <https://doi.org/10.1093/oso/9780198738831.003.0006> ('Blocking as Counter-Speech').

50 de Silva (n 16).

51 Langton, 'Beyond Belief' (n 49).

52 Ibid 72.

53 Some actors may also exploit shifts in norms to act on permissibility facts enacted through sex-based vilification regardless of their own views or feelings.

54 de Silva (n 16).

55 For example, a recent study found that women members of Parliament in Sweden feel significantly constrained in their speech as a result of abusive speech they are subjected to online. In particular, they 'avoid certain topics that are perceived as generating a great deal of online abuse', with one participant noting that discussions around gender equality and migration 'trigger the trolls rather quickly': Josefina Erikson, Sandra Håkansson and Cecilia Josefsson, 'Three Dimensions of Gendered Online Abuse: Analyzing Swedish MPs' Experiences of Social Media' (2023) 21(3) *Perspectives on Politics* 896, 906 <https://doi.org/10.1017/S1537592721002048>.

work, among other things, all also play significant roles.[56] Moreover, the extent to which sex-based vilification subordinates and silences women, or will do so over time, in fact, causally speaking, is an empirical question that cannot be precisely answered.[57] However, women's material oppression in patriarchal societies may be seen to flow from their systemic constitution and reconstitution as subordinate and silenced in those societies.[58] And speech acts of sex-based vilification contribute to – in that they are speech acts of – that constitution.[59]

That sex-based vilification constitutes and causes harm in these ways has significance beyond its impacts on individual women and women as a group, to the character of democracy. Silencing harms are particularly relevant here. Sex-based vilification functions, and is often intended, to exclude women from full democratic participation. This is especially true of sex-based vilification that occurs in spaces in which core political communication also occurs, including platforms. For many women, as for many others, platforms are now key loci of public discourse and engagement in public life. In liberal democracies, platforms are specifically also key loci of women's participation in democratic processes of the kind that, according to liberal arguments, serve to legitimate exercises of public power over and affecting them.[60] In turn, women's presence in and engagement within those spaces, or lack thereof, pertains to democracy itself.

Sex-based vilification occurring on platforms thus warrants careful and urgent consideration in liberal democracies for multiple compelling reasons. It necessitates equally effective regulatory responses to address its harms.

III PLATFORMED SEX-BASED VILIFICATION

Sex-based vilification occurring on platforms has particular gravity because of the way in which patriarchal and platform power intersect. In this Part, we combine our functional conceptualisation of the harms of such speech extrapolated in Part II with Fuchs' typology for analysing 'three faces' of global corporate power in

56 de Silva and Simpson (n 41) 499.
57 de Silva (n 16) 1022.
58 Ibid.
59 Ibid.
60 Arguments around the legitimating function of communication in democracies have been developed primarily by Ronald Dworkin and James Weinstein: see, eg, Ronald Dworkin, *Taking Rights Seriously* (Harvard University Press, 1978) 15–26; Ashutosh Bhagwat and James Weinstein, 'Freedom of Expression and Democracy' in Adrienne Stone and Frederick Schauer (eds), *The Oxford Handbook of Freedom of Speech* (Oxford University Press, 2021) 82, 98–103. See also Frederick Schauer, *Free Speech: A Philosophical Enquiry* (Cambridge University Press, 1982) 19–20. As to the importance of *online* communication to the relevant democratic processes, see, eg, James Weinstein, 'Cyber Harassment and Free Speech: Drawing the Line Online' in Susan J Brison and Katharine Gelber (eds), *Free Speech in the Digital Age* (Oxford University Press, 2019) 52, 53; Robert C Post, 'Privacy, Speech, and the Digital Imagination' in Susan J Brison and Katharine Gelber (eds), *Free Speech in the Digital Age* (Oxford University Press, 2019) 104, 110.

political processes.[61] We use Fuchs' framework to illuminate how platform power works to exacerbate the relevant subordination and silencing harms to women by platforming – that is, by amplifying, accommodating and authorising – sex-based vilification occurring on platforms. We return to Fuchs's typology in Part V, where we consider the role that the discursive power of platforms plays in the auspicing of platformed sex-based vilification through their construction of themselves as legitimate self-governors with respect to such speech.

A Three Faces of Platform Power

Prompted by the need to explain the political power of business actors in international relations and implications for democratic governance, Fuchs proposes differentiating three dimensions of corporate power: 'instrumental', 'structural' and 'discursive'.[62] As she argues, each dimension varies 'regarding the sources of power on which they draw, as well as the channels through which power is exercised'.[63] This approach recognises that power can be based on control over either material or ideational resources or both.

Power can also be exercised in ways that are either or both actor-specific and structural. Actor-specific power refers to power exercised or experienced by individuals as part of transactions or interactions between them. Structural power refers to power exercised through institutions and systems like corporations, shareholder capitalism or the digital speech economy, as well as the doctrines and practices of law.

By recognising the multi-dimensionality of corporate power, this approach allows for a more nuanced examination of the contingency of platform power than could be achieved with reference to framings that are entirely sceptical about corporate political power or those that propose 'undifferentiated claims of corporations ruling the world'.[64] Fuchs's three faces of power enable analysis of how different sources and channels of platform power can interact to shore each other up and create what seems to be an unassailable wall of power. Equally, it enables us to identify where potential for contestation of one aspect of platform power may have broader, more transformative ramifications.

Fuchs' discussion of instrumental power focuses on the direct influence by corporate special interests on political decision-makers through material economic resources, such as paid lobbying and campaign financial contributions.[65] However, for the purposes of our analysis, we adapt this approach to focus on the instrumental power of platforms towards those 'below', meaning the people who participate in self-presentation and communication on platforms. From this perspective, instrumental power refers to platforms' direct control of the computational

61 Fuchs, 'Theorizing Power' (n 15). See also John Mikler, *The Political Power of Global Corporations* (Polity Press, 2018).

62 Mikler (n 61) develops and applies Fuchs' approach in Fuchs, 'Theorizing Power' (n 15).

63 Fuchs, 'Theorizing Power' (n 15) 77.

64 Doris Fuchs, *Business Power in Global Governance* (Lynne Rienner Publishers, 2007) 4 <https://doi.org/10.1515/9781685853716> ('*Business Power*').

65 Fuchs, 'Theorizing Power' (n 15) 80. See also Mikler (n 61) 35–40.

affordances and infrastructures through which individuals communicate with family, friends and the wider world.

Fuchs uses structural power to refer to the economic processes that global corporations control, typically through the concentration of market power, where only one or a few corporations dominate options for buying or selling particular products and services in a particular domain.[66] Historically, the concentration of market power, a form of material economic power, was also seen as a dangerous step toward concentration of political or ideational power.[67] Similar concerns are resurging in current attempts to enforce competition law against big tech companies.[68] Many commentators have also noted the dangers that are now evident in relation to the structural power of platforms as significant 'infrastructures' for social, economic and political discourses.[69] Likewise, platforms' instrumental power in controlling the affordances and infrastructures that form the core of their value proposition, of augmenting and maximising opportunities for presentation, networking and connection, is dependent on a sufficient mass of users. Platforms' instrumental power and their (infra)structural power, in the sense of their dominance over our speech economies, are thus closely interrelated.

Finally, discursive power refers to the power of corporate actors to (re)create ideas as 'truth' and claim alignment between their interests and the interests of states and individuals.[70] Discursive power is closely connected to instrumental and structural power as it acts to legitimate the exercise of these other aspects of corporate power typically through the claim that the corporate actor must be trusted to self-regulate in the public interest.[71] This does not mean that discursive power renders instrumental and structural power intrinsically (or normatively) legitimate in fact.[72] Instead, it facilitates the pragmatic or sociological acceptance of legitimacy in which corporate self-regulation, 'comes to be institutionalized as "normal," or at the very least "tolerable"'.[73] Thus, self-regulation 'endures and institutionally re-embeds itself rather than having to be continually asserted'.[74] Discursive power thus reinforces instrumental and structural power. Or, to put it

66 Mikler (n 61) 20.

67 For example, recognition of this danger motivated the creation of antitrust laws in the 1890s in the United States, to counter the power of railroad and other monopolies created in the Gilded Age. See Tim Wu, *The Curse of Bigness: Antitrust in the New Gilded Age* (Columbia Global Reports, 2018) 10.

68 Ibid; Lina M Khan, 'Sources of Tech Platform Power' (2018) 2(2) *Georgetown Law Technology Review* 325; 'Digital Platform Services Inquiry 2020–2025', *ACCC: Australian Competition and Consumer Commission* (Web Page, 2024) <https://www.accc.gov.au/inquiries-and-consultations/digital-platform-services-inquiry-2020-25>.

69 Our references to (infra)structural power draws on the work of Jimena Valdez, as discussed at nn 118–19 below and accompanying text.

70 Mikler (n 61) 45.

71 See, eg, Ronen Shamir, 'Capitalism, Governance, and Authority: The Case of Corporate Social Responsibility' (2010) 6 *Annual Review of Law and Social Science* 531 <https://doi.org/10.1146/annurev-lawsocsci-102209-153000>.

72 Mikler (n 61) 45.

73 Ibid 48–9, quoting Stephen Wilks, *The Political Power of the Business Corporation* (Edward Elgar Publishing, 2013) 177. Regarding the sociological legitimacy of Meta's oversight board, see above n 32.

74 Mikler (n 61) 49. Note here the conceptual and practical analogousness of platforms' discursive power and the constitutive force of sex-based vilification, as discussed in Part II.

the other way around, to change the way platforms exercise their instrumental power through the affordances and infrastructures at the heart of their business models and to challenge platforms' structural power that their instrumental power enables (and vice versa), advocates for reform must also challenge and contest platforms' discursive power.

B Amplification of Sex-Based Vilification on Platforms

Attributes of online communication that commonly occur on platforms amplify the prevalence and severity of sex-based vilification. This is the first way in which sex-based vilification occurring on platforms may be said to be platformed.

In a widely cited article, psychologist John Suler argues that what he terms the 'online disinhibition effect',[75] means that 'people say and do things in cyberspace that they [would not] ordinarily say and do in the face-to-face world.'[76] In the absence of social cues, online participants are more likely to act without pause or restraint.[77] Users 'may [also] progress more steadily and quickly towards deeper expressions of … disinhibition that avert[s] social norms'.[78] When manifesting as 'toxic disinhibition',[79] these dynamics amplify the regularity and severity of harmful online conduct. They may lead to 'rude language, harsh criticisms, anger, hatred, even threats'.[80] Other authors have argued similarly. Danielle Keats Citron observes that '[i]f you cut data, it doesn't bleed. So [you are] at liberty to do anything you want to people who are not people but merely images'.[81] Martha Nussbaum notes in relation to the objectification of women online that, in contrast to the 'self-enclosed, self-nourishing world' of the internet, '[i]n daily life, there are some barriers to a woman's conversion from "whole and usual" into a mere set of stigmatized organs'.[82] These observations are consistent with the virulence that is characteristic of sex-based vilification directed at and about women online, including on platforms, as evident in the testimonies of targets[83] and some perpetrators.[84]

---

75 John Suler, 'The Online Disinhibition Effect' (2004) 7(3) *CyberPsychology and Behavior* 321 <https://doi.org/10.1089/1094931041291295>.

76 Ibid 321.

77 Ibid 322. Social cues here may include 'a frown, a shaking head, a sigh, a bored expression, and many other subtle and not so subtle signs of disapproval or indifference'.

78 Ibid 323.

79 Ibid 321.

80 Ibid.

81 Citron, *Hate Crimes* (n 3) 59, quoting Teresa Wiltz, 'Cyberspace Shields Hateful Bloggers: Death of Rapper's Mother Elicits Venomous Insults', *The Journal Gazette* (Fort Wayne, 17 November 2007).

82 Martha C Nussbaum, 'Objectification and Internet Misogyny' in Saul Levmore and Martha C Nussbaum (eds), *The Offensive Internet: Speech, Privacy, and Reputation* (Harvard University Press, 2011) 68, 74 <https://doi.org/10.2307/j.ctvjf9zc8.7>.

83 For example, *Jezebel*'s Anna North observes that 'she has been called evil, ugly, and sexless online, but she [does not] experience that kind of abuse offline': Citron, *Hate Crimes* (n 3) 59. See generally at 57–9.

84 For example, online message board posters who engaged in a vicious campaign of cyber harassment against two female law students from Yale University wrote under various, largely nonsensical, pseudonyms. One poster, whose real identity was ultimately uncovered, observed of his behaviour: 'I didn't mean to say anything bad … What I said about her was absolutely terrible, and I deserve to have my life ruined. I said something really stupid on the … internet, I typed for literally, like, 12 seconds':

Networking and its related effects of group polarisation are further attributes of online communication that commonly occur on platforms and that amplify sex-based vilification. Just as online communication makes it easier for socially benign or beneficial networks to form, so too does it more easily facilitate convergence on falsehoods, nefarious cyber mobs and antisocial networked conduct.[85] In this context, group dynamics may polarise mob speakers and other actors online such that the regularity and toxicity of their harmful conduct is intensified. Platform users may articulate and act on more extreme views to be accepted as credible within a group and because being part of a group increases their confidence and sense of belonging.[86] As Citron notes, when cyber mob members 'engage in an ever-escalating competition to destroy victims', the 'crowd-sourced nature of the destruction disperses feelings of culpability for some [members]', but 'the effects [on victims] are concentrated to the extreme'.[87] As with disinhibition, networking and group polarisation effects are evident in sex-based vilification directed at women on various platforms, especially as part of campaigns of cyber harassment.[88]

Related to networks and group polarisation are information cascades.[89] These, too, are relatively common on platforms and may amplify some of the sex-based vilification occurring on platforms. More broadly than the term suggests, 'information' cascades can sometimes be more akin to *dis*information cascades or what Tobias Rose-Stockwell describes as 'outrage cascades – viral explosions of moral judgment and disgust'.[90] They can occur when a person observes the behaviours, beliefs or emotions of others and, despite contradictions in those behaviours, beliefs or emotions that she herself knows to exist, engages in those same behaviours or takes on those same beliefs or emotions. Crucially, '[o]nce an information cascade gains momentum, it can be difficult to stop'.[91] Though information cascades also occur in the real world, the ease with which online communications can reach vast numbers of people both synchronously and asynchronously means that the effects of information cascades are heightened online.[92] As we discuss below, this ability to reach very large crowds of people both synchronously and asynchronously is a key affordance of platforms. (Dis)information and outrage cascades are characteristic of

---

Citron, *Hate Crimes* (n 3) 57. A man that operated various online fora dedicated to 'publish[ing] nude pictures of young girls and women without their consent explained that his pseudonymous character was merely "playing a game … "': at 59 (citations omitted).

85 Cass R Sunstein, 'Believing False Rumors' in Saul Levmore and Martha C Nussbaum (eds), *The Offensive Internet: Speech, Privacy, and Reputation* (Harvard University Press, 2011) 91, 99. See also Citron, *Hate Crimes* (n 3) 61–2.

86 See, eg, Citron, *Hate Crimes* (n 3) 63.

87 Ibid 64. See generally at 64–5.

88 Emma Alice Jane describes misogynistic online commentators responding to a particular video of a cheerleader falling from a human pyramid as competing to produce the '"winning" entry in the shockability stakes': Jane (n 8) 561 (citations omitted).

89 Nussbaum (n 82) 92–3, 95–6.

90 Tobias Rose-Stockwell, 'Facebook's Problems Can Be Solved with Design', *Quartz* (Web Page, 30 April 2018) <https://qz.com/1264547/facebooks-problems-can-be-solved-with-design>, quoted in Luke Munn, 'Angry by Design: Toxic Communication and Technical Architectures' (2020) 7 *Humanities and Social Sciences Communications* 53:1–11, 5 <https://doi.org/10.1057/s41599-020-00550-7>.

91 Citron, *Hate Crimes* (n 3) 66.

92 Ibid.

the sex-based vilification that women experience on platforms as part of campaigns of cyber harassment against them.[93]

As commonly occurring attributes of online communication on platforms, disinhibition, networks and group polarisation, and information cascades are interrelated and reinforcing. Each attribute benefits from the *overall* polarised communicative and normative environments partly constituted by the other attributes, while itself contributing to those environments in ways that are directly supportive and nurturing of the other attributes.[94] Accordingly, the problem is not so much that platforms amplify (or, as discussed below, accommodate or authorise) sex-based vilification through 'filter bubbles' or 'echo chambers', the evidence for which may be lacking, as Axel Bruns argues.[95] The problem is the overall polarisation, or platform users' holding of and acting on hyperpartisan and extremist views, which then (re)constitute conditions that are felicitous for the amplification (and accommodation and authorisation) of sex-based vilification occurring on platforms.[96]

While platforms are not 'the root cause of such developments', they 'do provide a forum for … [polarisation], by enabling … extremists to amplify each other's voices and coordinate their activities more efficiently'.[97] Indeed, the communicative phenomena described above are exacerbated by the very affordances and infrastructures that *characterise* platforms and make them attractive, because of the distinctive nature of the communication that platforms enable. That platforms amplify sex-based vilification is an emanation of their control over the affordances and infrastructures that form the core of their value proposition. Thus, the amplification of sex-based vilification occurring on platforms may be seen to be an aspect of the platforms' exercise of their instrumental power or their power to directly influence how those who use their services communicate with one another.

In their proposed definition of social media, Caleb T Carr and Rebecca A Hayes summarise the key features of social media as enabling communication through 'disentrained, persistent channels' and 'masspersonal' communication.[98] This highlights that platform users do not need to be present at the same time online to communicate with one another or to maintain the perception of persistent social interaction and discourse.[99] It also highlights that platforms allow users to broadcast

---

93 See, eg, the instances referred to at nn 4–7 above.

94 Note that the communicative attributes we discuss here are also not exclusive. There may be others that contribute in salient ways.

95 Axel Bruns, *Are Filter Bubbles Real?* (Polity Press, 2019) <https://doi.org/10.14763/2019.4.1426>.

96 Ibid 105.

97 Ibid 106. Bruns argues, for instance, that 'many of the well-known causes of polarisation still persist: for example, socioeconomic inequalities, citizen disenfranchisement, and inflammatory propaganda': at 105–6. We would argue that the structural oppression of some groups by others also fundamentally contributes.

98 Carr and Hayes (n 24) 50–2. Carr and Hayes identify five key features of social media: they are internet-based; they facilitate communication by disentrained and persistent channels; they provide a perception of interactivity (though not necessarily actual interactivity); the value of using them is user-generated; and they allow for 'masspersonal' communication.

99 See ibid 50, citing Joseph B Walther, 'Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction' (1996) 23(1) *Communication Research* 3 <https://doi.org/10.1177/009365096023001001>, where Carr and Hayes describe 'channel disentrainment' as:

messages to very large audiences, yet each user feels they are sending and receiving messages in a personalised way.[100] Together, these features help create and heighten the opportunity for the psychological effects of online disinhibition on individual users while maximising the potential for collective dynamics such as group polarisation and (dis)information and outrage cascades. Harmful communicative conduct can thus seem individual and intimate yet also be replicated, and felt, by many users simultaneously.[101]

Platforms' instrumental power to amplify harmful speech is particularly pernicious where the algorithmic recommender systems on which platforms heavily rely are designed to privilege material that promotes prolonged engagement through speech stoking conflict and contestation. Luke Munn highlights how Facebook's 'News Feed' feature, for instance, in order to maximise users' engagement and time spent on the platform, exposes users not only to relational content to do with their 'friends' but also to content engendering division, shock and outrage.[102] That is, *by design*, Facebook's imperative to increase engagement includes outrage expression and it thus 'works to reduce the barrier[s] to outrage expression'.[103] Munn argues that '[a]t its worst … Facebook's Feed stimulates the user with outrage-inducing content while also enabling its seamless sharing, allowing such content to rapidly proliferate across the network'.[104]

> [C]ommunication facilitated by a particular channel in which the user participates when he or she can commit to participating, as opposed to face-to-face communication, when both members of the communication dyad need to be committed at the same time … Its root, entrainment, comes from the organizational behavior and natural sciences literature and means to adjust one's pace or cycle to match that of another … thus, disentrained means that this adjustment is unnecessary.

See also Deborah Ancona and Chee-Leong Chong, 'Entrainment: Pace, Cycle, and Rhythm' in Barry M Staw and LL Cummings (eds), *Research in Organizational Behavior* (JAI Press, 1996) vol 18, 251.

100 Carr and Hayes (n 24) 52.

101 Ibid 54.

102 Munn (n 90). See Amnesty International, *The Social Atrocity: Meta and the Right to Remedy for the Rohingya* (Report, 29 September 2022) 42 <https://www.amnesty.org/en/documents/asa16/5933/2022/en/> ('*The Social Atrocity*'), where a Facebook employee explains in the 'Facebook Papers' released by whistleblower Frances Haugen:

> [E]ngagement doesn't necessarily mean that a user actually wants to see more of something. One of our biggest signals we use to provide more similar content is comments and … a comment that you hate a thing can be seen as a positive signal leading content to get outsized distribution. People game this in various ways, posting ever more outrageous things to get comments and reactions that our algorithms interpret as signs we should allow things to go viral …

103 Munn (n 90) 5. Munn notes that 'Facebook has admitted that hate speech is a problem and has redesigned the Feed dozens of times since its debut in an effort to curtail this problem'. But, he argues, the core logic of engagement remains baked into the design of the Feed at a deep level.

104 Ibid. For example, Amnesty International found that Facebook's engagement-based algorithmic systems assisted in proactively amplifying and promoting content that incited hatred against the Rohingya in Myanmar, which in turn contributed to the 2017 massacre of Rohingya in that country: *The Social Atrocity* (n 102); 'Myanmar: Facebook's Systems Promoted Violence Against Rohingya', *Amnesty International* (Web Page, 28 September 2022) <https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/> ('Facebook's Systems Promoted Violence'). Similarly, the introduction of new features in Facebook's systems including the angry emoji and weighting of different reactions contributed to promoting ethnic violence in Ethiopia in 2020: Amnesty International, *'A Death Sentence for My Father': Meta's Contribution to Human Rights*

Thus, the amplification of sex-based vilification on platforms may be seen to be directly related to how platforms '[encourage] forms of hate speech that are spontaneous in the sense of being instant responses, gut reactions, unconsidered judgments, off-the-cuff remarks, unfiltered commentary, and first thoughts'.[105] The very aspects of platforms that allow freer expression in desirable ways make it '*cheap* to slur someone' or to otherwise harm them, including through conduct constituting sex-based vilification.[106] And each of the characteristics of online communication that Suler identifies as contributing to disinhibition[107] in ways that might amplify the prevalence and severity of sex-based vilification are readily *facilitated by* the design features that are central to platforms' business models and that set platforms apart from other fora.

C Accommodation and Authorisation of Sex-Based Vilification on Platforms

The affordances and infrastructures at the core of platforms' business models also appear to enable communicative environments that render sex-based vilification particularly cumulative and reinforcing. This is the second way in which sex-based vilification occurring on platforms may be said to be platformed.

In Part II, we discussed how permissibility facts enacted through sex-based vilification may be accommodated as correct play or, conversely, challenged or undone.[108] Online disinhibition, networks and group polarisation effects, and (dis) information or outrage cascades appear to contribute to the accommodation of sex-based vilification as correct play by and for many platform users. As also discussed in Part II, this in turn increases the likelihood that such speech causes harms to women in fact. Additionally, these dynamics problematise the *potential* for platform users to challenge or undo the harms of such speech. They render sex-based vilification occurring on platforms less vulnerable to influences, particularly counter-narratives, that could mitigate its subordination and silencing of women. For example, that online disinhibition may partly be reflective of the internet, including platforms, as 'a self-enclosed, self-nourishing world that is remarkably resistant to the reality outside',[109] suggests that much sex-based

---

Abuses in Northern Ethiopia (Report, 31 October 2021) 42–4 <https://www.amnesty.org/en/documents/afr25/7292/2023/en/>. Both reports relied on internal Facebook papers released by Facebook whistleblower Frances Haugen who linked Facebook's algorithmic systems to hate speech and ethnic violence: Karen Hao, 'The Facebook Whistleblower Says Its Algorithms Are Dangerous. Here's Why', *MIT Technology Review* (online, 5 October 2021) <https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/>.

105 Alexander Brown, 'What Is So Special about Online (As Compared to Offline) Hate Speech?' (2018) 18(3) *Ethnicities* 297, 304 <https://doi.org/10.1177/1468796817709846>.

106 Saul Levmore and Martha C Nussbaum, 'Introduction' in Saul Levmore and Martha C Nussbaum (eds), *The Offensive Internet: Speech, Privacy, and Reputation* (Harvard University Press, 2011) 1, 2 <https://doi.org/10.4159/9780674058767-001> (emphasis added).

107 Namely, dissociative anonymity, invisibility, asynchronicity, solipsistic introjection, dissociative imagination, and minimisation of status and authority: Suler (n 75).

108 Langton, 'Beyond Belief' (n 49) 83; Langton, 'How to Undo' (n 49); Langton, 'Blocking as Counter-Speech' (n 49).

109 Nussbaum (n 82) 74.

vilification occurring on platforms is accommodated rather than undone.[110] Group polarisation effects and (dis)information or outrage cascades also mean that like-minded individuals coalesce in networks that converge in their thinking and are thus especially accepting and supportive, as opposed to challenging, of fellow users' conduct.[111] In combination, these communicative phenomena may also partly explain the prolificity on platforms of hate groups, including 'incel' groups and other networks of users united by misogyny.[112]

If sex-based vilification is especially accommodated on platforms, it is also especially authorised on platforms, in the sense that accommodation authorises speakers and speech acts in aggregate and circular ways.[113] For instance, a Manosphere blogger may acquire authority by virtue of his subscribers celebrating, agreeing with or even omitting to question his pronouncements. The accommodation of sex-based vilification on platforms may thus be conceived of as conferring authority on such speech, such that its speakers may *successfully* or more *effectively* subordinate and silence women through their speech.[114] At least some speakers of sex-based vilification (beginning) with only the kind of covert authority described in Part II may in this way come to have substantive – if informal –[115] authority on platforms and may do more harm as a result.

The structural power of platforms as significant channels for self-presentation and communication is particularly relevant and problematic in accommodating and authorising sex-based vilification in these ways. Platforms now leverage a platform ecosystem in which people effectively *must* participate for many social, economic and political purposes. They are essential infrastructure not just in and for the high spaces of public political life but for everyday engagement. Given their market dominance, many individuals and groups have little choice but to participate in and on platforms, in order to have access and a voice in their local

---

110 This does not mean, however, that this speech has no consequences for women 'in the real world'. See above Part II. See also Katharine Gelber and Susan J Brison, 'Digital Dualism and the "Speech as Thought" Paradox' in Susan J Brison and Katharine Gelber (eds), *Free Speech in the Digital Age* (Oxford University Press, 2019) 12 <https://doi.org/10.1093/oso/9780190883591.003.0002> for a general critique of the idea that online speech is more like thought than action and thus harmless.

111 Sunstein (n 85) 93. Participants in information cascades also typically do not disclose knowledge they privately hold that might bring into question the behaviours, beliefs, or emotions they are receiving and passing on, so 'the judgment of group members will not reflect the overall knowledge, or the aggregate knowledge, of those within the group – even if the information held by individual members, if actually revealed and aggregated, would produce a better and quite different conclusion'.

112 See, eg, Citron, *Hate Crimes* (n 3) 62. See also Michael Flood, 'Men's Rights: A Collection of Accessible Critiques', *XY Online* (Web Page, 20 February 2015) <https://xyonline.net/content/mens-rights-collection-accessible-critiques>.

113 Langton, 'Beyond Belief' (n 49); Rae Langton, 'Accommodating Authority' (John Locke Lecture, University of Oxford, 29 April 2015) <https://media.philosophy.ox.ac.uk/assets/mp3_file/0019/37126/Lecture_1.mp3>.

114 In Austinian speech act terms, accommodation may be a 'felicity condition' of some such speech. John L Austin, *How To Do Things with Words: The William James Lectures Delivered at Harvard University in 1955* (Oxford University Press, 1975) 14–15.

115 See Rae Langton, 'Speech Acts and Unspeakable Acts' (1993) 22(4) *Philosophy and Public Affairs* 293, 329. Substantive authority may be formal (for example, in the case of a legislator) or it may be informal or customary (for example, in the case of a parent in relation to her adult child).

communities and political affairs, market a product or business (via advertising and selling), participate in the workforce, or maintain relationships with family and friends.[116] Indeed, platforms' structural power in this sense may be self-fulfilling to such an extent that it even displaces (and replaces) trust in other, established 'off-platform' social processes that previously performed the same functions.[117]

Writing of Uber and workforce participation, Jimena Valdez conceptualises this aspect of platform power as '*infra*structural power',[118] a term we find useful in elucidating the close relationship between platforms' instrumental power as discussed in the previous section and their structural power as discussed in this section. As Valdez writes, the course of platforms' infrastructural power 'is tied to their normal operation – that is, the provision of their services – and stems from the specific position these firms occupy in the economy' as mediators between producers and consumers, enabling an ecosystem on which producers and consumers depend,[119] as well as, in the case of social media platforms, social and political networks. As noted, platforms' infrastructural power is evidenced by the virtual impossibility for individual participants to exit them without compromising their family, friendship, business and civic networks, as well as significant personal and relational data such as photographs, conversations and details about people, places and events, known as 'switching costs'.[120]

Thus, platforms' infrastructural power renders them highly significant in the creation of both public and private discourses. Whatever model for speech they promote is not just amplified but becomes constitutive of the nature of what speech is regarded as acceptable per se. Julie Cohen refers to this as the 'emergent limbic media system' and draws on the metaphor of the neurological system to point out the way in which networks created and governed by online platforms create a sort of 'digital unconscious' of what is sayable and unsayable, thinkable and unthinkable.[121]

Platforms as constitutive of discursive and social norms are also 'limbic' in another sense. As Antonia Lyons and colleagues argue, 'limbic platform capitalism' operates by using data to train algorithmic models that stimulate

---

116 See Cory Doctorow, *The Internet Con: How to Seize the Means of Computation* (Verso, 2023) 7; Salomé Viljoen, 'A Relational Theory of Data Governance' (2021) 131(2) *Yale Law Journal* 573, 616.

117 Where 'process' refers to 'an ordered assemblage of conventions, rules, steps, institutions, norms, etc, guiding action, activity, and practice toward the attainment of specific ends': Laufer and Nissenbaum (n 12) 14.

118 Jimena Valdez, 'The Politics of Uber: Infrastructural Power in the United States and Europe' (2023) 17(1) *Regulation and Governance* 177 <https://doi.org/10.1111/rego.12456> (emphasis added). This conception derives from a broader literature on digital media platforms in critical infrastructure studies to which Valdez refers.

119 Ibid 177. Valdez differentiates infrastructural from both instrumental and structural power, explicitly citing a Fuchs type analysis. We differ from Valdez for the purposes of our analysis in this article insofar as we conceptualise infrastructural power as an extension of Fuchs' conception of structural power, rather than as an alternative or additional dimension of power.

120 Doctorow (n 116) 7.

121 Julie E Cohen, 'The Emergent Limbic Media System' in Mireille Hildebrandt and Kieron O'Hara (eds), *Life and the Law in the Era of Data-Driven Agency* (Edward Elgar Publishing, 2020) 60, 61 citing Mireille Hildebrandt, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar Publishing, 2015). See also Julie E Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press, 2019) 76–7.

neuropsychological reward processes in the brain relating to pleasure, mood, habit and attention.[122] They focus on the way social media affordances use sensory pleasure to 'stimulate and sustain users' attention'[123] in order to '"tune" the flow of content that users encounter'[124] and 'nudge' consumption of products and services advertised, including especially harmful and dangerously addictive consumption of alcohol and gambling.[125] While Lyons and colleagues write about the manipulation of algorithmic processes towards pleasure in a commercial context, it is also true that platforms can, and are often designed to, stimulate the limbic system and prolong engagement by amplifying shock, anger, and hate in a discursive context, as we discussed in the previous section in relation to the amplification of sex-based vilification. Thus, platforms' relevant affordances may exploit negative 'triggers' and be *addictive*[126] and 'enable anger to spread *contagiously*'.[127] As Max Fisher and Amanda Taub write, the 'incentive structures and social cues of algorithm-driven social media sites' may even create 'real-world extremists' by training users over time to '*arrive* at hate speech'.[128] Similarly, Munn argues in relation to the amplification of 'outrage-inducing content' on Facebook's News Feed that 'in increasing the prevalence of such content and making it easier to share, it becomes normalized. Outrage retains its ability to provoke engagement, but in many ways becomes an established aspect of the environment.'[129]

In the case of oppressive speech, like sex-based vilification, what is sayable and thinkable (that is, what is normalised) constitutes the lived realities of target group members. Where platforms' infrastructural power facilitates their 'tuning' of content flows and manipulating of limbic processes en masse, the platforming of sex-based vilification through its accommodation and authorisation on platforms may be seen to be an aspect of that power. Importantly, it may be seen to be an aspect of platforms' infrastructural power that subordinates and silences women in systemic and material ways.

---

122 Antonia C Lyons et al, 'Limbic Platform Capitalism: Understanding the Contemporary Marketing of Health-Demoting Products on Social Media' (2023) 31(3) *Addiction Research and Theory* 178, 179 <https://doi.org/10.1080/16066359.2022.2124976>; Christine Parker et al, 'Addressing the Accountability Gap: Gambling Advertising and Social Media Platform Responsibilities' (2023) 32(4) *Addiction Research and Theory* 312, 313 <https://doi.org/10.1080/16066359.2023.2269852>.

123 Lyons et al (n 122) 180.

124 Ibid 179. Nicholas Carah, Daniel Angus and Jean Burgess, 'Tuning Machines: An Approach to Exploring How Instagram's Machine Vision Operates on and through Digital Media's Participatory Visual Cultures' (2023) 37(1) *Cultural Studies* 20 <https://doi.org/10.1080/09502386.2022.2042578>.

125 Lyons et al (n 122) 180.

126 Munn (n 90) 2, citing Paul Lewis, '"Our Minds Can Be Hijacked": The Tech Insiders Who Fear a Smartphone Dystopia', *The Guardian* (online, 6 October 2017) <https://www.theguardian.com/technology/2017/oct/05/smartphone-addiction-silicon-valley-dystopia>.

127 Munn (n 90) 2, citing Rui Fan, Ke Xu and Jichang Zhao, 'Higher Contagion and Weaker Ties Mean Anger Spreads Faster than Joy in Social Media', *arxiv* (Online Paper, 12 August 2016) <http://arxiv.org/abs/1608.03656> (emphasis added).

128 Munn (n 90) 2, citing Max Fisher and Amanda Taub, 'How Everyday Social Media Users Become Real-World Extremists', *The New York Times* (online, 25 April 2018) <https://www.nytimes.com/2018/04/25/world/asia/facebook-extremism.html> (emphasis added).

129 Munn (n 90) 5. See also *The Social Atrocity* (n 102) and 'Facebook's Systems Promoted Violence' (n 104).

IV REGULATORY RESPONSES TO PLATFORMED SEX-BASED VILIFICATION

We suggested in the previous section that sex-based vilification is platform*ed*, as a matter of platforms' instrumental and infrastructural power, in ways that shape how women are spoken to and about and thus treated in patriarchal societies. In this Part, we consider existing responses to platformed sex-based vilification, with a focus on platform accountability. We show that there is a 'gap' in the law with respect to platforms' accountability for such speech. To the extent that the regulatory landscape constructs platforms as responsible for platformed sex-based vilification at all, it privileges self-regulatory measures. These measures are in turn inadequate and inappropriate to address the relevant systemic subordination and silencing harms to women.

A Existing Legal and Other Regulatory Responses

1 Overview

Speech occurring on platforms would normally be subject to existing anti-vilification laws.[130] Apart from a few notable exceptions,[131] however, there are no laws against vilification on the basis of sex (or gender).[132] Some extant laws, including on threatening conduct, harassment, sexual harassment and obscenity, for example, may incidentally capture some speech that constitutes sex-based vilification. But as discussed below, these are neither directed nor effective at addressing the causal and constitutive harms of such speech *as vilification*. In any

130 See, eg, the definition of acts occurring 'otherwise than in private' for the purposes of Australia's federal racial vilification law: *Racial Discrimination Act 1975* (Cth) s 18C ('*Cth Discrimination Act*'). Cf n 133 below regarding the Canadian position.

131 See, eg, *Anti-Discrimination Act 1992* (NT) s 20A; *Criminal Code*, RSC 1985, c C-46, ss 318(4), 319(1)–(2), (7); *Promotion of Equality and Prevention of Unfair Discrimination Act 2000* (South Africa) s 10. See also *Strafgesetzbuch* [Criminal Code] (Germany) § 130 [tr authors]. In June 2020, the Cologne Upper Regional Court held that 'sections of the population' for the purposes of that provision includes women: 'German Hate Speech Laws Also Cover Misogynist Abuse, Court Rules'*, Deutsche Welle* [German Wave] (online, 15 June 2020), archived at <https://perma.cc/23J4-RE4Z>.

132 It is unclear why this is the case. In contrast, vilification on the basis of other ascriptive characteristics, including, for example, disability, gender identity, HIV/AIDS status, intersex status, race, religion or sexuality is unlawful under international law and in many domestic jurisdictions. Each of those categories of vilification is prohibited in varying forms in one or more Australian jurisdictions: *Cth Discrimination Act* (n 130) s 18C; *Criminal Code Act 1995* (Cth) ss 80.2A–80.2B, 80.2D ('*Cth Criminal Code*'); *ACT Discrimination Act* (n 20) s 67A; *ACT Criminal Code* (n 20) s 750; *Anti-Discrimination Act 1977* (NSW) ss 20C, 38S, 49ZT, 49ZXB; *NSW Crimes Act* (n 20) s 93Z; *QLD Anti-Discrimination Act* (n 20) ss 124A, 131A; *Civil Liability Act 1936* (SA) s 73; *Racial Vilification Act 1996* (SA) s 4; *Anti-Discrimination Act 1998* (Tas) s 19; *Racial and Religious Tolerance Act 2001* (Vic) ss 7–8, 24–5; *Criminal Code Act Compilation Act 1913* (WA) ss 77–80. For examples of categories of vilification prohibited in other domestic jurisdictions: see Alex Brown, *Hate Speech Law: A Philosophical Examination* (Routledge, 2015) ch 2. Racial and religious vilification is also prohibited under international law: *International Convention on the Elimination of All Forms of Racial Discrimination*, opened for signature 21 December 1965, 660 UNTS 195 (entered into force 4 January 1969) arts 1, 4; *International Covenant on Civil and Political Rights*, opened for signature 19 December 1966, 999 UNTS 171 (entered into force 23 March 1976) art 20.

case, though individuals may be held civilly or criminally liable for their online speech under existing anti-vilification laws, platforms would not automatically be captured as speakers themselves, such that they would be liable for the vilifying content they host (or platform). Certainly, in the few jurisdictions in which sex-based (and/or gender-based) vilification laws do exist, there appears to have been no case law to date to suggest that platforms would be held legally accountable for such speech.[133]

There have been some recent, significant developments in laws directed at regulating content on platforms in Australia, the UK and the EU, in the forms of the Australian and UK Online Safety Acts and the *Digital Services Act*.[134] These are discussed in the next section. Outside of these, sections 474.33–474.34 of the *Criminal Code Act 1995* (Cth)[135] are rare examples of provisions explicitly directed at regulating content on platforms. Pursuant to those provisions, internet service providers or corporations providing a 'content service' or 'hosting service', which includes platforms,[136] may be held criminally liable if they knowingly host 'abhorrent violent material that records or streams abhorrent violent conduct that has occurred, or is occurring, in Australia' and fail to report it to the Australian Federal Police within a reasonable amount of time.[137] Platforms may also be held criminally liable if they do not 'ensure the expeditious removal' of abhorrent violent material.[138] Similar provisions exist in the UK and under European law.[139] These kinds of laws may capture recordings of actual rape or other extreme violence against women.[140] It is unclear whether they would capture more normalised violence or imagery that constitutes sex-based vilification but that 'merely' depicts, rather than records, violence against women, as much mainstream heterosexual pornography available online is typically deemed to do, for example.[141]

---

133 In Canada, ongoing consultation on potential laws expressly targeted at online hate suggest that online communications may be seen as distinct from speech covered by Canada's existing anti-vilification laws: see above n 131. See Garth Davies and Sarah Negrin, 'Regulating Online Hate Will Have Unintended, but Predictable, Consequences', *The Conversation* (online, 3 June 2022) <https://theconversation.com/regulating-online-hate-will-have-unintended-but-predictable-consequences-182724> for some background.

134 See above n 18.

135 *Cth Criminal Code* (n 132) ss 474.33–474.34.

136 See ibid s 474.30(a) (definition of 'content service' and 'hosting service').

137 Ibid s 474.33.

138 Ibid s 474.34.

139 See, eg, Nicolas P Suzor, *Lawless: The Secret Rules That Govern Our Digital Lives* (Cambridge University Press, 2019) 49–51.

140 See *Cth Criminal Code* (n 132) ss 474.31 (definition of 'abhorrent violent material'), 474.32 (definition of 'abhorrent violent conduct').

141 Of course, much mainstream heterosexual pornography available online does, in fact, record actual violence done to actual women, whether or not it is treated accordingly for the purposes of law. The e-Safety Commissioner's 2022–23 annual report notes that only three notices were issued (to overseas service providers) regarding 'abhorrent violent material': Australian Communications and Media Authority and eSafety Commissioner, *Annual Report 2022–23* (Report, 2023) <https://www.esafety.gov.au/sites/default/files/2023-10/ACMA-and-eSafety-Commissioner-annual-report-2022-23.pdf>. It is unclear what the particular content of the material was in relation to notices issued.

Platforms may also 'opt in' to non-binding regulatory regimes guiding, but not enforcing, particular content moderation standards. An example of this is the EU Code of Conduct on Countering Illegal Hate Speech Online ('EU Code of Conduct'),[142] which applies to member states and platforms that have signed up to it. The EU Code of Conduct is voluntary and non-coercive and relies instead on dialogue between platforms and the European Commission.[143] Regular evaluations of the EU Code of Conduct in practice are published detailing user reporting, removal rates, time of assessment, feedback to users and transparency across platforms, as well as grounds on which users reported hateful content,[144] but it is up to platforms themselves to provide this information. As the EU Code of Conduct's name suggests, its concern is to counter hate speech that the EU has sought to see member states criminalise under domestic law. The European Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law,[145] with reference to which the EU Code of Conduct defines the obligations of member states and platforms, refers only to public incitements of violence or hatred against target group members 'defined by reference to race, colour, religion, descent or national or ethnic origin'.[146] Vilification on the basis of sex (or gender) is not captured at all.

2 Recent Developments

(a) Australia

In Australia, 'social media services', may now be fined pursuant to the *Online Safety Act 2021* (Cth) ('*AU Safety Act*')[147] for failure to comply with a notice from the Australian e-Safety Commissioner to remove specific instances of non-consensual pornography,[148] which is a kind of sex-based vilification and 'cyber-abuse' targeted at Australian adults.[149] 'Cyber-Abuse' is defined to mean material

142 'Code of Conduct on Countering Illegal Hate Speech Online' (Code, European Commission, 30 June 2016) <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>.

143 As of November 2022, 36 organisations from 21 Member States sent notifications to the IT companies taking part in the Code of Conduct: Didier Reynders, 'Countering Illegal Hate Speech Online: 7th Evaluation of the Code of Conduct' (Factsheet, European Commission, November 2022) <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> ('7th Evaluation of the Code of Conduct').

144 Ibid.

145 *Council Framework Decision 2008/913/JHA of 28 November 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law* [2008] OJ L 328/55.

146 Ibid art 1(1)(a).

147 *AU Safety Act* (n 18) s 13 (definition of 'social media services').

148 Ibid ss 77, 80.

149 *AU Safety Act* (n 18) ss 88, 91. Some material depicting 'abhorrent violent conduct' is also covered in pt 8, subject to the eSafety Commissioner's satisfaction that the availability of the material online is 'likely to cause significant harm to the Australia community': see, eg, at s 95. These provisions thus capture an even narrower subset of platformed sex-based vilification than do the criminal provisions discussed above: see above nn 135–41 and accompanying text.

that 'an ordinary reasonable person would conclude … is likely … intended to have an effect of causing serious harm to a particular Australian adult' and 'an ordinary reasonable person in the position of an Australian adult would regard … as being, in all the circumstances, menacing, harassing or offensive'.[150] 'Serious harm' includes temporary or permanent physical and psychological harm.[151]

Under the *AU Safety Act*, the eSafety Commissioner is also charged with administering an 'Online Content Scheme' that covers certain content that is captured by Australia's National Classification Scheme.[152] The National Classification Scheme takes as its basis obscenity laws' 'rationale of protecting the community or sections of the community from exposure to publications that offend community standards'.[153] For example, the Classification Board's content classification decisions must take into account 'the standards of morality, decency and propriety generally accepted by reasonable adults'.[154] The National Classification Scheme and thus the Online Content Scheme incidentally capture some sex-based vilification, especially 'depictions that condone or incite violence, particularly sexual violence' and 'the portrayal of persons in a demeaning manner'.[155]

The recent Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023 (Cth) ('Misinformation and Disinformation Bill')[156] proposes to give the Australian Communications and Media Authority various powers with respect to the dissemination to Australian users of content constituting 'misinformation' and 'disinformation' on 'digital services', which includes platforms.[157] The Misinformation and Disinformation Bill in its current form captures vilifying speech that is 'reasonably likely to cause or contribute to serious harm',[158] with 'harm' being defined to include 'hatred against a group in Australian society' on the basis of various protected characteristics, including gender.[159] This threshold of, effectively, inciting or 'contributing to' 'serious' hatred appears to be a higher threshold of harm than what is covered by many anti-vilification laws,[160] and it is unclear how it would be applied in practice. The Misinformation and Disinformation Bill is in any case likely to undergo significant change before it is passed, if at all, due to concerns around its impact on free expression, particularly religious speech.[161]

150 *AU Safety Act* (n 18) s 7(1).
151 Ibid s 5 (definition of 'serious harm').
152 Ibid pt 9.
153 Des Butler and Sharon Rodrick, *Australian Media Law* (Thomson Reuters, 6th ed, 2021) 740.
154 *Classification (Publications, Films and Computer Games) Act 1995* (Cth) s 11(a).
155 *National Classification Code* (Cth) s 1(d)(i)–(ii).
156 Misinformation and Disinformation Bill (n 20).
157 Ibid sch 1 s 4.
158 Ibid sch 1 s 7.
159 Ibid sch 1 s 2 (definition of 'harm').
160 In the Australian context, see, eg, *Cth Discrimination Act* (n 130) s 18C, which refers to acts that are 'reasonably likely' to 'offend, insult, humiliate or intimidate'.
161 Josh Taylor, 'Labor to Overhaul Misinformation Bill after Objections over Freedom of Speech', *The Guardian* (online, 13 November 2023) <https://www.theguardian.com/australia-news/2023/nov/13/labor-misinformation-bill-objections-freedom-of-speech-religious-freedom>.

(b) UK

In the UK, the *Online Safety Act 2023* (UK) ('*UK Safety Act*') applies to any platform that has, is capable of having, or aims to have users in the UK.[162] Platforms may additionally be captured as 'Category 1' services by delegated legislation.[163] These are anticipated to be the largest service providers with the most users[164] and will have more onerous obligations.

The *UK Safety Act* imposes relatively wide-ranging requirements on platforms in relation to monitoring and removing 'illegal' content.[165] This includes an obligation to 'swiftly take down' illegal content where a platform is made aware of it.[166] As flagged above, sex-based vilification is not currently prohibited by law in the UK. Platforms are also required to 'prevent individuals from encountering *priority* illegal content'[167] and 'minimise the length of time for which any priority illegal content is present'.[168] 'Priority illegal content' includes 'content that amounts to' certain criminal offences,[169] including offences relating to harassment, stalking, threatening or inciting violence,[170] possession of 'extreme' pornographic images,[171] non-consensual pornography,[172] and family violence.[173] Some such content may, in some circumstances, encompass communicative conduct that constitutes sex-based vilification and that is thus incidentally captured. Notably, the provisions on priority illegal content do directly target some hate speech, namely, incitements to hatred on the basis of race, religion and sexual orientation.[174] However, no direct protections are afforded to women against vilification on the basis of sex (or gender).[175]

During its passage into law, the Online Safety Bill 2022 (UK)[176] was widely criticised for its failure to account for women's distinct vulnerabilities and the systemic harms they regularly face online.[177] A proposal to encompass violence against women and girls as priority illegal content and address it specifically was

162 *UK Safety Act* (n 18) ss 4(2), (5), (6).
163 Ibid sch 11 s 1(1). The delegated legislation is still pending.
164 Ibid.
165 See generally ibid ss 9–10.
166 Ibid s 10(3)(b).
167 Ibid s 10(2)(a) (emphasis added).
168 Ibid s 10(3)(a).
169 Ibid s 59(10)(c).
170 Ibid sch 7 ss 4, 6–10.
171 Ibid sch 7 s 29.
172 Ibid sch 7 s 31.
173 Ibid sch 7 s 11.
174 Ibid sch 7 s 5. This reflects existing anti-vilification laws in England and Wales. See, eg, *Public Order Act 1986* (UK) pts 3, 3A.
175 The *UK Safety Act* (n 18) also captures some communicative conduct amounting to racially or religiously aggravated harassment or public order offences (see, eg, ibid sch 7 s 9(a)), but it has no such protections in place for their sex- or gender-based equivalents.
176 Online Safety Bill 2022 (UK).
177 See, eg, 'Women and Girls Failed by Government's Online Safety Bill', *End Violence against Women* (Web Page, 17 March 2022) <https://www.endviolenceagainstwomen.org.uk/women-girls-failed-governments-online-safety-bill/>.

ultimately defeated.[178] Category 1 platforms are, however, tasked with additional requirements directed at 'user empowerment' with respect to discriminatory content, including abusive content on the basis of sex or content that incites hatred on the basis of sex.[179] This appears to relate to the provision to users of tools that would enable them to have better control over the content they encounter, for example by blocking and muting other users whose content they do not wish to see.

The Office of Communications ('OFCOM'), the administrative entity in charge of the *UK Safety Act*, is tasked with developing relevant codes of practice and guidance relating to the Act, and a platform will be treated as complying with a duty under the Act if it complies with corresponding recommendations in a code of practice.[180] The codes themselves are non-binding.[181] As part of this broader obligation, OFCOM must produce guidance in relation to 'content and activity … which disproportionately affects women and girls'.[182] Among other things, the guidance may 'contain advice and examples of best practice for assessing risks of harm to women and girls … and for reducing such risks'.[183]

Platforms are also required under the *UK Safety Act* to operate complaints procedures in relation to the content described above.[184] Importantly, platforms' 'duties' under the Act do not just include obligations in relation to the operation of platforms, for example, content moderation in response to communicative conduct that occurs on platforms. They also include obligations to design platforms so that risks and harms associated with captured content, including as set out above, are minimised.[185] It is as yet unclear what this will mean in practice.

(c) EU

In Europe, the *Digital Services Act* now imposes a range of obligations on intermediaries. These obligations are particularly onerous for platforms (meaning VLOPs), which are required, among other things, to monitor and respond to 'illegal' content and content giving rise to certain 'systemic risks'. Platforms are obligated to respond to authorities' orders to take down illegal content,[186] as well as take down illegal content that they become aware of.[187] They are also required to notify authorities of 'suspicions' of criminal offences involving threats to life and safety.[188] They must put mechanisms in place to enable individuals and other

178 United Kingdom, *Parliamentary Debates*, House of Commons, 12 July 2022, vol 718, col 173 (Alex Davies-Jones); United Kingdom, *Parliamentary Debates*, House of Commons, 28 June 2022, vol 717, col 650 (Kirsty Blackman).
179 *UK Safety Act* (n 18) ss 14, 15, 16(4)(c), 16(5)(a).
180 See generally ibid s 49.
181 Ibid s 50.
182 Ibid s 54(1). OFCOM must also consult on the guidance with the Commissioner for Victims and Witnesses, the Domestic Abuse Commissioner, and others as appropriate: at s 54(3)).
183 Ibid s 54(2)(a).
184 Ibid s 21.
185 See, eg, ibid ss 9(5)(h), 14(5)(g).
186 *Digital Services Act* (n 18) art 9.
187 Ibid art 6.
188 Ibid art 18.

entities to notify them of illegal content[189] and provide reasons to those persons explaining their (the platform's) response to such content.[190] Platforms must additionally put in place internal complaints handling processes to deal with complaints regarding their responses,[191] suspend repeat offenders,[192] and report on their content moderation outcomes with respect to their obligations under the *Digital Services Act*.[193] Finally, platforms are required to monitor 'systemic risks … stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services'.[194] The relevant systemic risks may relate to the dissemination of illegal content, actual or foreseeable negative effects to fundamental rights – including to human dignity, private life, protection of personal data, free expression and non-discrimination – and actual or foreseeable negative effects relating to 'gender-based violence'.[195] Platforms are also required to 'put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified … with particular consideration to the impacts of such measures on fundamental rights'.[196] The development of voluntary and self-regulatory standards and their adoption is incentivised as a potentially relevant consideration in determining whether these risk assessment and mitigation obligations have been complied with.[197] As with the *UK Safety Act*, it may be that some conduct that is distinctly prohibited by other criminal laws and that has communicative elements constituting sex-based vilification – for example, credible threats of sexual violence or the dissemination of non-consensual pornography – is covered by these aspects of the *Digital Services Act*. Overall, however, and notwithstanding the *Digital Services Act*'s references to systemic risks, fundamental rights, and risk identification and mitigation, there is no evidence that these broad concepts would be interpreted to extend to cover sex-based vilification in the absence of explicit changes to underlying laws regionally or in the relevant domestic jurisdictions.[198]

B The Gap in the Law

Notwithstanding that they incidentally capture some speech constituting platformed sex-based vilification, the existing measures described above represent a 'gap' in the law with respect to platform liability for such speech and its harms to women. Most obviously, they capture only a small subset of the most serious

189 Ibid art 16.
190 Ibid art 17.
191 Ibid art 20. See also at art 21 regarding out of court dispute settlement.
192 Ibid art 23.
193 Ibid art 24.
194 Ibid art 34.
195 Ibid arts 34(1)(a)–(b), (d).
196 Ibid art 35.
197 See ibid Preamble para 104. See also at arts 34(2)(b)–(c), 35(2)(b)–(c).
198 Ibid arts 34, 35. See also Marta Maroni, '"Mediated Transparency": The Digital Services Act and the Legitimisation of Platform Power' in Maarten Hillebrandt, Päivi Leino-Sandberg and Ida Koivisto (eds), *(In)visible European Government: Critical Approaches to Transparency as an Ideal and a Practice* (Routledge, 2024) 305.

conduct constituting platformed sex-based vilification. For example, obligations on platforms to respond to content that constitutes non-consensual pornography or that rely on obscenity law standards generally only address sex-based vilification to the extent that such speech is sexualised or otherwise explicit. However, not all sex-based vilification is sexualised or explicit.

Platform governance standards that rely on existing criminal laws, as with the *UK Safety Act* and the *Digital Services Act*, are similarly limited. Criminal harassment offences, for instance, typically require perpetrators to have engaged in a 'course' of conduct,[199] whereas sex-based vilification may manifest as 'one-off' utterances. As with obligations relating to non-consensual pornography and cyber abuse or bullying, as present in the *AU Safety Act*, such criminal laws also require that the offending conduct be directed at individual, identifiable women.[200] They do nothing to capture platformed sex-based vilification that is about women as a group. That is, platformed sex-based vilification includes some of, but is much broader than, the communicative conduct captured by existing laws and regulatory measures that touch on platform governance in key jurisdictions.

Additionally, because the measures described above are not directed at platformed sex-based vilification as vilification, they are not addressed to the relevant systemic harms of such speech. They conceptualise contemptuous, sex-based platformed speech directed at women as giving rise to harms, if any, that are disparate, situational and seemingly unrelated. They therein obfuscate the shared, cumulative and reinforcing functions of such speech, as well as the *reasons* we may wish to see it regulated. Obscenity law standards, for instance, on which the *AU Safety Act* Online Safety Scheme is reliant,[201] are simultaneously too narrow and too wide to appropriately and adequately address the relevant harms. Laws regulating 'obscene' or 'indecent' speech do so not on the basis that such speech harms women in and by subordinating and silencing them for being women, but on the basis that such speech is unpalatable, immoral or otherwise not in accordance with prevailing community standards.[202] Legal formulations and judgments of what is 'good' and 'bad' explicit content, when made to reflect patriarchal community norms, may also actively disserve women.[203]

---

199 In the Australian context, for example, state and territory laws criminalise stalking. Those offences often target behaviour amounting to harassment, which typically requires that perpetrators have engaged in a course of relevant conduct. See, eg, *Crimes Act 1958* (Vic) s 21A ('*Vic Crimes Act*').

200 Ibid.

201 See above nn 152–154 and accompanying text.

202 See *R v Hicklin* (1867–68) LR 3 QB 360, which sets out the common law test for obscenity: at 371 (Cockburn CJ). See also *R v Close* [1948] VLR 445 for an instance of the move away from 'obscenity' to 'indecency'. See, eg, *Summary Offences Act 1966* (Vic) s 17(1)(b) for a legislative example in the Australian context.

203 That is, some sex-based speech that systemically subordinates and silences women is not captured by obscenity laws, whereas some speech is captured by obscenity laws notwithstanding that it does not harm as platformed sex-based vilification does. Sonya Sceats, for example, describes obscenity law as 'a conversation about morality [rather than harm] in which the participants (such as the publishers, prosecutors and judges) have overwhelmingly been men': Sonya Sceats, 'The Legal Concept of Obscenity: A Geneology' (2002) 16(1) *Australian Feminist Law Journal* 133, 143. Catharine MacKinnon argues that

Other standards encompassed by the laws described above, say, legal thresholds relating to criminal harassment and stalking, overlook the discriminatory or sex-based nature of platformed sex-based vilification and that its harms are systemic. They incidentally capture some speech that subordinates and silences women, but, again, they are not addressed to those harms as suffered by women as a group and *because* they are women.[204]

This gap in the law means that platform self-regulation through internal policies and practices constitutes the most prevalent and material category of extant response to platformed sex-based vilification within the broader regulatory landscape. As we show in the next parts, this undermines the regulation of platformed sex-based vilification in materials ways and leaves women at the mercy of platforms in seeking protection from such speech as they go about their lives online. In particular, it consolidates platform power with respect to platformed sex-based vilification and strengthens the intersection of patriarchy and platform power in auspicing sex-based vilification.

C Platform Self-Regulatory Responses

Many platforms have in place internal policies and content moderation conventions that form self-regulatory regimes. Some of these capture platformed sex-based vilification to a greater or lesser degree.[205] Overall, there is little to no transparency around how these internal policies work in practice, including how the categories of content they proscribe are interpreted or applied, either manually or through automated decision-making systems, and how they interact with business models that seek to optimise engaging content and thus amplify, accommodate and authorise sex-based vilification in the ways discussed. Recent reforms only go so far to remedy this. For instance, they require platforms to publish some metrics and explanatory information about their content moderation policies and reasons for specific content moderation decisions, but these metrics and information are

---

> [i]n practice, [obscenity laws] prohibit … depictions of sex that some men find offensive – that is, the public showing of sex that some men want to say they do not want other men to see. It takes the view that sex is dirty, women are dirty, and homosexuality is bad … It cares more about whether men blush than whether women bleed … Virtue and vice are its concerns; women and children are not.

Catharine A MacKinnon, 'Pornography's Empire' (Conference Paper, Commonwealth Law Conference, 16–20 April 1990), quoted in Regina Graycar and Jenny Morgan, *The Hidden Gender of Law* (Federation Press, 2nd ed, 2002) 405.

204 This is primarily because the conduct in question needs to be directed at individual, identifiable women for the relevant provisions to have force. See, eg, *Vic Crimes Act* (n 199) s 21A.

205 See, eg, 'Hateful Conduct Policy' (n 40); 'Community Standards', *Meta* (Web Page) <https://transparency.meta.com/en-gb/policies/community-standards> ('Meta Community Standards'); 'Safety and Civility', *TikTok* (Web Page, 17 April 2024) <https://www.tiktok.com/community-guidelines/en/safety-civility/>; 'Community Guidelines', *Privacy, Safety, and Policy Hub* (Web Page) <https://values.snap.com/en-GB/privacy/transparency/community-guidelines>; 'The X Rules', *X Help Center* (Web Page) <https://help.x.com/en/rules-and-policies/x-rules>; 'Hate Speech Policy', *YouTube Help* (Web Page) <https://support.google.com/youtube/answer/2801939?sjid=9445357372787287143-AP>; 'Community Guidelines', *Twitch* (Web Page) <https://safety.twitch.tv/s/article/Community-Guidelines?language=en AP>; 'Professional Community Policies', *LinkedIn* (Web Page) <https://www.linkedin.com/legal/professional-community-policies>.

largely decontextualised.[206] Moreover, in the context of the *Digital Services Act*, for example, Marta Martoni has suggested that these measures, alongside the obligations for platforms to put in place their own internal risk assessment and mitigation policies, operate more 'as a legitimising force for digital media platforms than as manoeuvres against the power structure of these platforms'.[207] Certainly, as Martoni argues, some legal and regulatory approaches may institutionalise platform self-regulation by encouraging and incentivising voluntary standards, self-regulatory codes of conduct, and the notion that platforms themselves are the appropriate authorities to decide whether to take down unlawful content, as long as they report on their reasons for decision.[208] This is a problem because platform self-regulation is, prima facie and in practice, neither adequate nor appropriate to address the systemic harms to women of platformed sex-based vilification.

Consider in this respect Facebook's Community Standards,[209] which incorporate a hodgepodge of guidelines on prohibited and restricted content. Its Community Standards on 'hateful conduct' ('Hateful Conduct Policy')[210] defines prohibited hateful conduct as 'direct attacks against people – rather than concepts or institutions' on the basis of a range of protected characteristics, including sex.[211] This includes 'dehumanising speech, allegations of serious immorality or criminality, and slurs … harmful stereotypes … serious insults, expressions of contempt or disgust, swearing and calls for exclusion or segregation'.[212] 'Harmful stereotypes' are defined as 'dehumanising comparisons that have historically been used to attack, intimidate or exclude specific groups, and that are often linked with offline violence'.[213] Also proscribed are examples of language deemed to be hate speech or not, in decreasing severity from 'Tier 1' to 'Tier 2'.[214]

The Hateful Conduct Policy defines sex-based hate speech expressively, contrary to our functional theory of the harms of such speech.[215] It also includes

206 See above nn 190 and 193 and accompanying text in relation to the *Digital Services Act* (n 18). The *AU Safety Act* and *UK Safety Act* also contain reporting obligations: see, eg, *AU Safety Act* (n 18) s 183(2); *UK Safety Act* (n 18) s 77. See also Paddy Leerssen, 'Outside the Black Box: From Algorithmic Transparency to Platform Observability in the Digital Services Act' (2024) 4(2) *Weizenbaum Journal of the Digital Society* 3.

207 Maroni (n 198) 305.

208 Ibid.

209 'Meta Community Standards' (n 205).

210 'Hateful Conduct Policy' (n 40). See also above nn 35–7 and accompanying text. As with the FOB, it is unclear what role the new policy on 'hateful conduct' will play in light of the accompanying changes to Facebook's content moderation practices.

211 'Hateful Conduct Policy' (n 40). Sex, sexual orientation and gender identity are protected characteristics but not gender.

212 Ibid.

213 Ibid.

214 Ibid. A separate Community Standard on 'violence and incitement' governs speech that 'incites or facilitates violence [including against a group of people on the basis of a protected characteristic] and credible threats to public or personal safety': 'Violence and Incitement', *Meta* (Web Page) <https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/>. And other Community Standards cover 'adult sexual exploitation', 'bullying and harassment', 'violent and graphic content', and a range of other conduct: 'Meta Community Standards' (n 205).

215 See de Silva (n 16) 1021–2 for an extrapolation of the advantages of a functional approach.

some questionable exemptions that could relate to sex/gender-based speech directed at or about women, the most obvious of these being for 'certain gender-based cursing in a romantic break-up context' and 'content arguing for gender-based limitations of military, law enforcement and teaching jobs'.[216] In previous iterations, it has also been too prescriptive and thus too narrow in relation to some speech. Sex-based vilification, including platformed sex-based vilification, often serves to objectify women in the absence of explicit references or comparisons to women as 'objects', 'household objects' or 'property', which are all terms that were expressly prohibited and served as standards of reference in the Hateful Conduct Policy up until 7 January 2025.[217] Finally, the Hateful Conduct Policy is 'sex neutral', in that it covers speech directed at or about women and men. While this is consistent with existing sex-based vilification laws, it nevertheless fails to reflect that sexed, contemptuous speech directed at and about men does not, and cannot, systemically harm them in the same ways that sex-based vilification harms women in patriarchal societies.[218] This can be especially problematic in the context of content moderation standards that are prescriptively and inflexibly framed. For example, terms such as 'whore' and 'slut', were also expressly prohibited by the Hateful Conduct Policy up until 7 January 2025.[219] However, these mean different things for, and have different impacts on, women and men, as female and male sexuality are judged differently in patriarchal societies.

Platform self-regulation is also inadequate and inappropriate to address the harms of platformed sex-based vilification in practice. For example, the literature on Facebook's content moderation failures to date in relation to harmful, discriminatory speech is vast. The leaking of the platform's content moderation training manuals in 2017[220] also provides some insights. As Amy Binns notes, on the basis of the manuals, Facebook 'allow[s] … content most users would find abhorrent', including graphically detailed credible threats of violence and calls to violence against women.[221] Even direct threats to particular women might be allowed if no particulars as to the target's movements are provided alongside.[222] Feminist campaigners have further highlighted that Facebook regularly and disproportionately applies its content moderation rules, including the Hateful

---

216 'Hateful Conduct Policy' (n 40).

217 See the section titled 'Tier 1' in the version of the 'Hateful Conduct Policy' (n 40) accessed by the authors at 29 February 2024. Note that references to these terms have been removed from the updated policy as part of the suite of changes to Facebook's content moderation policies and practices in early 2025. See above nn 36–8 and accompanying text.

218 See below n 223 and accompanying text.

219 See above n 217. See the section titled 'Tier 2' in the version of the 'Hateful Conduct Policy' (n 40) accessed by the authors at 29 February 2024.

220 See, eg, Nick Hopkins, 'Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence', *The Guardian* (online, 22 May 2017) <https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence>.

221 Amy Binns, 'Facebook's Moderation Rules Prove It's OK with Being a Hostile Place for Women', *The Conversation* (online, 23 May 2017) <https://theconversation.com/facebooks-moderation-rules-prove-its-ok-with-being-a-hostile-place-for-women-78200>.

222 Ibid.

Conduct Policy, to remove phrases calling attention to male violence against women or other forms of patriarchal oppression.[223]

Self-regulation also embeds a logic of individualised 'responsibilisation' in which platforms' complicity in platformed sex-based vilification is obscured, and the problem of such speech and its solution are seen as located with individual platform users.[224] Platforms' internal content moderation policies and practices, like the Hateful Conduct Policy, tend to frame individual 'bad apple' perpetrators as creating and wholly responsible for platformed sex-based vilification, rather than addressing the pattern of affect promoted by the system. They frame other users, particularly women who are the targets of such speech, as having in hand the solution to the problem. This typically involves women users identifying, reporting and otherwise policing sex-based vilification occurring on platforms, as well as policing themselves, to attempt to protect themselves from harm.[225]

Finally, self-regulation often relies on faith in a 'technological fix',[226] being the idea that algorithmic models can be trained to address deep social problems, while allowing platforms to avoid investing in effective 'human' interventions through the proper framing and implementation of policy, as well as the safer design of their offerings.[227] Platforms use their discursive power to claim the perfectibility of their computational processes, including that the relevant algorithms can ultimately be taught to identify and demote harmful content, especially if, again, individuals take

---

223 See, eg, Simon van Zuylen-Wood, '"Men Are Scum": Inside Facebook's War on Hate Speech', *Vanity Fair* (online, 26 February 2019) <https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech>. This is notwithstanding a relevant exemption for such speech apparently contained in its internal content moderation guidance. A recent FOB case deals with this issue, with the FOB overturning a decision by Facebook to remove two posts condemning male violence against women and recommending 'that Meta include the exception for allowing content that condemns or raises awareness of gender-based violence in the public language of the Hate Speech policy, as well as update its internal guidance to reviewers to ensure such posts are not mistakenly removed': 'Oversight Board Overturns Meta's Decisions in the Violence against Women Cases', *Oversight Board* (Web Page, 12 July 2023) <https://oversightboard.com/news/1664046764100847-oversight-board-overturns-meta-s-decisions-in-the-violence-against-women-cases/>; 'Violence against Women', *Oversight Board* (Web Page, 12 July 2023) <https://www.oversightboard.com/decision/IG-H3138H6S>. See also 'Call for Women's Protest in Cuba', *Oversight Board* (Web Page, 3 October 2023) <https://www.oversightboard.com/decision/IG-RH16OBG3/>.

224 For a discussion of how individualised 'responsibilisation' can be used to obscure organisational responsibility for harm in other areas of regulatory governance, see Garry C Gray, 'The Responsibilization Strategy of Health and Safety: Neo-Liberalism and the Reconfiguration of Individual Responsibility for Risk' (2009) 49(3) *British Journal of Criminology* 326 <https://doi.org/10.1093/bjc/azp004>.

225 See Rosalie Gillett, Zahra Stardust and Jean Burgess, 'Safety for Whom? Investigating How Platforms Frame and Perform Safety and Harm Interventions' (2022) 8(4) *Social Media and Society* 1 <https://doi.org/10.1177/20563051221144315>.

226 Christian Katzenbach, '"AI Will Fix This": The Technical, Discursive, and Political Turn to AI in Governing Communication' (2021) 8(2) *Big Data and Society* 1, 1 <https://doi.org/10.1177/20539517211046182>.

227 See, eg, Douek (n 32) 12, quoting Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement', *Facebook* (Blog Post, 15 November 2018) <https://perma.cc/C7P3-DLYT> for Douek's discussion regarding the use of AI in Facebook content moderation as being inappropriate in relation to hate speech (as well as bullying and harassment). She notes that Mark Zuckerberg has claimed that AI can quickly and proactively identify harmful content and is 'the single most important improvement in enforcing [Facebook's] policies'.

responsibility for identifying and marking that content.[228] In reality, these processes are replete with biases and errors that re-enact rather than mitigate discriminatory harms, including to women, as discussed in Part III.[229]

V AUSPICING BY PLATFORMS OF PLATFORMED SEX-BASED VILIFICATION

The privileging of self-regulatory action in the platform regulatory landscape and the resulting harmful outcomes for women are related to platforms' framing of themselves as legitimate arbiters of the treatment of the sex-based vilification they platform. In this section, we consider how platforms mobilise their discursive power to this end and the significance of this.

As discussed in Part III, affordances and infrastructures at the core of platforms' business models, through which platforms exercise their instrumental and infrastructural power, underly their platforming of sex-based vilification. Attempts to mitigate the harms of platformed sex-based vilification through external regulation, of either the operation or design of platforms, thus directly threaten platform power. And platforms' corporate and profit imperatives conflict with, and drive, platforms' responses to platformed sex-based vilification.

Just as platforms make claims as to their economic, social and political value, they mobilise their discursive power to 'construct themselves as legitimate self-governors'.[230] This places them in a 'privileged' position, albeit one that is often 'contested' and contingent, from which they may seek to maintain control or influence over the conditions for their own regulation and governance.[231] Platforms may resist regulation, shape it to their own purposes or comply with conditions only of their choosing. This may manifest, for instance, as efforts to keep computational processes non-transparent (and thus non-accountable), frame individuals (and thus not platforms) as responsible for managing their own speech or avoiding speech they find undesirable, or justify extant or proposed internal content moderation policies and practices (and thus avoid stricter regulation by external administrators).[232] The individualised nature of users' feeds or experiences and platforms' power to keep confidential and non-transparent the algorithmic systems that create these makes the extent and nature of platforms' amplification of sex-based vilification (as discussed above) less 'observable' to civil society and

228 See above nn 224–5 and accompanying text.

229 See, eg, Smartparenting Staff, 'Facebook Took Down a Beautiful Photo of a Breastfeeding Angelica Panganiban and We Are Puzzled', *spin.ph* (online, 12 October 2023) <https://www.spin.ph/life/people/facebook-took-down-a-beautiful-photo-of-a-breastfeeding-angelica-panganiban-and-we-are-puzzled-a2749-20231012>.

230 Mikler (n 61) 20.

231 Fuchs, *Business Power* (n 64) 4.

232 As discussed, these are key failings of the self-regulatory model. See Part IV(C) above.

public authorities, thus stymying opportunities for regulatory accountability.[233] Platforms can choose to increase transparency in the relevant computational processes to allow for greater oversight by other actors, including users and public authorities, however, they typically jealously guard this power by claiming they can be trusted to have appropriate policies in place and to self-enforce these.

Relatedly, platforms may appeal to seemingly universal, overriding values to minimise regulatory oversight or shun it altogether. The platform 'playbook' in this regard frequently includes referencing 'free speech' as a core part of their value proposition to be protected from external interference. This particular claim can be especially powerful in liberal democratic jurisdictions in shielding platforms from legal liability for the publication of discriminatory and otherwise harmful speech, including platformed sex-based vilification.[234] Josh Cowls et al argue that '"constitutional metaphors" as metaphorical allusions to concepts, institutions, or practices of constitutional democracy', are also used by platforms (and by others in relation to platforms) in ways that privilege platform self-regulation.[235] That is, '[c]onstitutional metaphors in platform governance … establish a novel connection between old concepts and new practices, with the effect of legitimating private institutions through association with public governance institutions'.[236]

The FOB is in many ways a paradigmatic example. Meta describes the FOB as 'apply[ing] … content standards in a way that protects freedom of expression and other global human rights standards' by 'providing an independent check on Meta's content moderation … [and] making binding decisions on the most challenging content issues'.[237] However, as Douek notes, the FOB 'cannot be expected to offer … procedural recourse or error correction in anything but the smallest fraction of … cases'.[238] And though it is in some ways legally and structurally independent from Facebook,[239] the FOB is not as independent from Meta as would be, say, a court enforcing state laws. It represents something much closer to self-regulation.

From Facebook's perspective, 'key amongst the reasons for establishing the FOB is the desire to find a way of legitimizing the power that Facebook exercises over its users and the public sphere'.[240] And 'Facebook's establishment of … [the

233 See, eg, Parker et al (n 122) 312. On platform 'observability', see Bernhard Rieder and Jeanette Hofmann, 'Towards Platform Observability' (2020) 9(4) *Internet Policy Review* 1 <https://doi.org/10.14763/2020.4.1535>.

234 See, eg, the history of the *Communications Act of 1934*, 47 USC § 230 (1934), which provides platforms and internet service providers with immunity from liability for content provided by users in the US: Suzor (n 139) 43–58. In relation to platforms like Pornhub, see, eg, Joseph Hogan, 'Section 230 Was Created to Shield Sites Like Pornhub. It Might Be Killed for the Same Reason', *Medium* (Blog, 30 December 2020) <https://medium.com/retro-report/section-230-was-created-to-shield-sites-like-pornhub-against-lawsuits-75cf0a11bbfc>.

235 Cowls et al (n 34) 2451.

236 Ibid 2450.

237 *Oversight Board* (Website, 2025) <https://www.oversightboard.com/>.

238 Douek (n 32) 5–6.

239 See, eg, Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129(8) *Yale Law Journal* 2418.

240 Douek (n 32) 7–8. See also Robert Gorwa, 'The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content' (2019) 8(2) *Internet Policy Review* 1 <https://doi.org/10.14763/2019.2.1407>;

FOB] makes it not merely a passive recipient of regulatory mandates but a proactive player in the design of the future of Internet governance',[241] assisting the company in particular with 'staving off or guiding more extensive government regulation'.[242] Cowls et al argue that the commonly deployed constitutional metaphor of the FOB as 'Facebook's Supreme Court' ascribes to it a 'quasi-sovereign' role that corresponds to Facebook's *assertion* of such a role.[243] This is significant because 'the use of "supreme court" as a descriptor for the [F]OB … inevitably brings with it the weighty social, cultural, and political capital that attaches to supreme courts, particularly in the United States'.[244] And, 'mapping common knowledge of the US Supreme Court onto the blank canvas of the [F]OB may bolster its image and ultimately … enhance the legitimacy of the Board and of Facebook itself'.[245] In other words, the FOB – its establishment, operation and promotion – is, materially, if not primarily, an exercise by Facebook of its discursive power; to validate its instrumental and infrastructural power over users and society and, relatedly, influence and ultimately minimise the regulatory oversight and enforcement to which it is subjected.

The establishment of the FOB, though particularly significant, is of course not the only time that Facebook has mobilised its discursive power for these ends. In 2020, Facebook published a 'white paper', which it called 'Charting a Way Forward: Online Content Regulation'.[246] In it, it framed the regulation of online content as 'a new governance challenge'[247] defined by 'four key questions'[248] of its own choosing and laid down 'principles for future regulators'.[249] The principles prioritised, among other things, incentives for platforms, 'freedom of expression', 'proportionality and necessity', and 'allow[ing] internet companies the flexibility to innovate'.[250] It also made explicit that the white paper represented 'the beginning of a conversation'[251] between 'policymakers and other stakeholders' as '[t]hese are challenging issues and sustainable solutions will require collaboration'.[252] Much more can be said about the substance of the paper, including, for example, the discursive power exercised through its prolific appeals to 'freedom of expression'[253]

Rotem Medzini, 'Enhanced Self-Regulation: The Case of Facebook's Content Governance' (2022) 24(10) *New Media and Society* 2227 <https://doi.org/10.1177/1461444821989352>.

241 Douek (n 32) 24.

242 Ibid 17. See also at 21–4.

243 Cowls et al (n 34) 2453. As to Facebook's role in the rise of the 'supreme court' metaphor, see at 2455–63.

244 Ibid 2452.

245 Ibid.

246 Monika Bickert, 'Charting a Way Forward: Online Content Regulation' (Policy Paper, Facebook, February 2020) <https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf>.

247 Ibid pt I.

248 Ibid pt II.

249 Ibid pt III.

250 Ibid 19–20.

251 Ibid pt IV.

252 Ibid 21.

253 See, eg, ibid 3–4, 7, 9, 11, 15–17, 19–20.

or declarations that platforms 'are intermediaries, not speakers'[254] and 'enforcement will always be imperfect'.[255] No doubt much can also be said of further examples, including from other platforms. The point is that platforms, like Facebook, can and do regularly engage their discursive power in explicit and implicit ways to attempt to shape the regulatory landscape to their will, in particular by attempting to privilege platform self-regulation.

To a greater or lesser extent, platforms' use of their discursive power to these ends is often also successful. This is not to say that there are no other factors contributing to states' reliance on self-regulatory (and co-regulatory) models in relation to platforms. For instance, states may feel they have neither the resources nor expertise to do away with such models entirely. However, it is clear that platforms' discursive power plays a material part in shaping the regulatory landscape to be as it is. A leading Australian media law textbook's consideration of 'the regulation of online content', for example, begins with a discussion of the Facebook white paper as representing the state of play, and it is then quoted and cited throughout.[256] The text even affords the paper quasi-directive, rather than merely consultative, status, suggesting that 'it remains to be seen whether forthcoming new laws will *comply with* the criteria proposed by Facebook'.[257] After Facebook's and other industry submissions on the *AU Safety Act* criticised the threshold of 'serious harm' under the Act for being too low,[258] the definition of 'serious harm to a person's mental health' was amended in Australia's Senate to exclude 'mere ordinary emotional reactions such as those of only distress, grief, fear or anger'.[259] Following the Charlie Hebdo attacks in France in 2015, the European Commission established the European Internet Forum, of which Meta is a member, to develop 'a joint, voluntary approach' for the detection and removal of 'online terrorism incitement and hate speech'.[260] What resulted was the voluntary EU Code of Conduct, despite European legislators having previously warned platforms that they would be subjected to onerous new laws in relation to the same.[261] Douek further notes that a recent report commissioned by the French government 'shows this dynamic

254 Ibid 7.

255 Ibid.

256 Butler and Rodrick (n 153) 770. To be clear, our observations here are not at all intended as criticisms of the authors of the text. Given Meta's dominance in the sector and its influence of the regulatory environment, as we seek to highlight, it is entirely expected that Facebook and its white paper would feature so prominently.

257 Ibid (emphasis added).

258 See, eg, 'Facebook's Response to Australian Government Consultation on a New Online Safety Act' (Policy Paper, Facebook, 19 February 2020) <https://about.fb.com/wp-content/uploads/2020/02/Facebook-response-to-consultation-new-Online-Safety-Act.pdf>. Facebook also suggested an alternative threshold of 'grossly offensive' rather than 'offensive' with respect to bullying and harassing content: at 24.

259 Butler and Rodrick (n 153) 783.

260 European Commission, 'EU Internet Forum: Bringing Together Governments, Europol and Technology Companies to Counter Terrorist Content and Hate Speech Online' (Press Release, 3 December 2015) <https://ec.europa.eu/commission/presscorner/detail/en/IP_15_6243>, cited in Danielle Keats Citron, 'Extremist Speech, Compelled Conformity, and Censorship Creep' (2018) 93(3) *Notre Dame Law Review* 1035, 1041.

261 Ibid 1040–4.

playing out, with Facebook's voluntary reforms influencing the model of regulation ultimately endorsed by the report's authors', who 'recommended a model that focused on "expanding and legitimizing" platform self-regulation based on "the progress made in the last 12 months by … Facebook"'.[262]

A regulatory environment in which platform power, especially discursive power, dictates the state of play is also suggested in the many quasi-self-regulatory aspects present in the *AU Safety Act* and *UK Safety Act*. In both Acts, mirroring the prevailing model adopted by platforms in their internal regimes, the burden falls on individual women who have been victimised to monitor for, identify and report even the little platformed sex-based vilification that is captured.[263] Both Acts also allow for supplementary industry codes of conduct for platforms to be administered and enforced by the relevant authorities. In each case, the development of the codes is to be driven by industry.[264] This emphasis on standard-setting by platforms, as the reference point from which regulators may impose or even imagine their own requirements, is consistent with the overall privileging of platform self-regulation that we argue is suggestive of platforms' active and impactful use of their discursive power in this space. It also further facilitates platforms' exercising their discursive power 'officially' and even more authoritatively, through state-sanctioned means, a strategy that is especially effective given the particular difficulties of regulatory enforcement in the digital space.

Thus, platforms wield their discursive power to attempt to legitimate and privilege inadequate and inappropriate self-regulatory measures with respect to platformed sex-based vilification, often successfully. These measures can in turn have the effect of discursively 'rubberstamping' ineffective or anti-feminist content moderation processes and outcomes in ways that re-enact platformed sex-based vilification's harms to women, as well as platforms' complicity in those harms. This may be taken to mean that platformed sex-based vilification – being speech that is amplified, accommodated and authorised on platforms – is also *auspiced by* platforms. This auspicing represents an additional manifestation or 'layer' of contempt for women, for which platforms currently are not but should be critiqued and held accountable.

Platforms' auspicing of platformed sex-based vilification is apparent in that their use of their discursive power to privilege self-regulation also undermines, in some material ways, potential for rendering platform power accountable for

---

262 Douek (n 32) 23, citing *Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France with a European Vision* (Mission Report, May 2019) <https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf> and showing that self-imposed platform reform and regulatory reform occur in dialogue with one another.

263 See, eg, *AU Safety Act* (n 18) ss 32 (regarding 'intimate images'), 36 (regarding 'cyber abuse material'); *UK Safety Act* (n 18) s 15 (regarding 'user empowerment'). For discussion, see above nn 224–5.

264 For example, the Online Content Scheme under the *AU Safety Act* states 'bodies or associations that the [eSafety] Commissioner is satisfied represent sections of the online industry should develop codes … that are to apply to participants in the respective sections of the industry in relation to their online activities': *AU Safety Act* (n 18) s 137(1). See generally at div 7. Similarly, OFCOM must consult with industry actors in developing its draft codes of practice as required under the *UK Safety Act*: *UK Safety Act* (n 18) s 41(6)(b).

platformed sex-based vilification. Self-regulation obfuscates and reinforces the harms of such speech by appearing to provide an internal governance solution that is not actually a solution, as discussed above,[265] thus reinforcing failures of existing laws to address platformed sex-based vilification. For example, in the latest evaluation of the EU Code of Conduct, hateful content on the basis of gender apparently comprised 4.1% of reported content as a whole.[266] However, gender (and sex) are not covered by the EU Code of Conduct.[267] And any reporting on the moderation of hate speech against women on the basis of gender (or sex) is merely incidental and has no regulatory basis in the EU Code of Conduct. Thus, to the extent that voluntary regimes such as the EU Code of Conduct rely on existing anti-vilification laws for their framing, they reinforce, rather than address, the sex-based gap in such laws discussed above. Even if the EU Code of Conduct and other such regimes reflected extended anti-vilification laws that included sex as a protected characteristic, relying on platforms to 'opt in', self-report and self-administer at least partly consolidates platform power with respect to platformed sex-based vilification. Because the EU Code of Conduct appears to provide a solution, however, it confuses the need and stymies future opportunities for more effective regulation.[268]

Platforms also auspice platformed sex-based vilification to the extent that their use of their discursive power to privilege self-regulation results in policy oversights, biased or overly narrow administration, or anti-feminist outcomes that further normalise sex-based vilification.[269] This danger was highlighted by the FOB, in the first case relating to misogynistic speech that it made available for public comment.[270] The case involved a decision by Facebook to remove a post containing a video in which a Spanish term meaning 'fag' was used. A term meaning 'bitch' was also used in the video, as part of the phrase 'son of a bitch', however, this latter term and the associated phrase were not emphasised by the FOB as a subject of their decision or as requiring comment. In English, 'fag' is, of course, a virulently homophobic slur. And words meaning 'fag' may, in the absence of further context, reasonably be characterised pursuant to the Hateful Conduct Policy as 'words that are inherently offensive and used as insulting labels' for a person or group of people on the basis of their sexual orientation.[271] It is significant, though, that Facebook emphasised the use of this word as the primary basis for the removal of the post.

This emphasis on the homophobic slur, as opposed to the misogynistic slur, is also prima facie inconsistent with the Hateful Conduct Policy. That is, even on

265 See above Part IV(C).

266 '7th Evaluation of the Code of Conduct' (n 143) 4.

267 See above nn 145–6 and accompanying text.

268 Indeed, this is exactly what happened: see above nn 263–4 and accompanying text.

269 In speech act terms, this is further accommodation of sex-based vilification: see above nn 48–50 and accompanying text.

270 'Oversight Board Overturns Facebook Decision in Columbia Protests Case', *Oversight Board* (Web Page, 27 September 2021) <https://www.oversightboard.com/news/223462609822963-oversight-board-overturns-facebook-decision-case-2021-010-fb-ua/>.

271 See the section titled 'Tier 3' in the version of the 'Hateful Conduct Policy' (n 40) accessed by the authors at 23 June 2021.

Facebook's own account, words meaning 'bitch' may, in the absence of further context, reasonably be characterised as 'words that are inherently offensive and used as insulting labels' for a woman or women on the basis of their sex. It is also difficult to see what jurisdiction- or language-specific context may mitigate such a finding in the context in which the term was used in the particular post. Moreover, 'content targeting … [women] on the basis of their … [sex] with … [c]ursing, defined as … [p]rofane terms or phrases with the intent to insult, including … bitch' was *expressly* proscribed by the version of the Hateful Conduct Policy in place at the time.[272] It is immaterial here that the term was used in the post as part of an attack on a man; men are often criticised or degraded on the basis of their relationships to (particular kinds of) women. Utterances in which terms meaning 'bitch' or similar appear are almost always about women, even where they are directed at men and even if they are said in jest, for example, and such utterances should accordingly be treated as directed at women. If the post in question was removed on the basis that it violates the Hateful Conduct Policy, it should thus have been removed with reference to its constituting both hate speech on the basis of sexual orientation (for its use of 'fag') *and* sex (for its use of 'bitch').

This brings to the fore the broader issue that sex-based vilification is, generally speaking, not only ubiquitous but also normalised. Such speech is often simultaneously overwhelming and invisible. The treatment of women as sexual objects or less than human, for instance, may be so central to social organisation in patriarchal societies that, unlike racist, homophobic or other hate speech, it is imperceptible as harm or as harm worth doing anything about. This dynamic may partly explain content moderation failures by platforms, like Facebook, to effectively identify and respond to platformed sex-based vilification and its harms to women in policy and practice.[273] But it also highlights that when platforms' use of their discursive power to privilege self-regulation results in their auspicing of platformed sex-based vilification, they are materially complicit in patriarchy itself.

VI CONCLUSIONS

We have argued that sex-based vilification occurring on platforms exists at the intersections of patriarchy and platform power and is platformed in two

272 Ibid.

273 Some progress does appear to be being made, however: see generally above n 223. See also Case 2023-014-IG-UA, in which the FOB overturned a decision by Meta to remove a video posted by a Cuban news platform on Instagram in which a woman is seen protesting against the Cuban government and calling for other women to join her on the streets. In contrast, she criticises men for failing to defend those who have been repressed and compares men to animals. The FOB found the speech in the video to include a qualified behavioural statement that, under the Hate Speech Community Standard, should be allowed: see above n 223. In Case 2023-015-FB-UA, a user appealed a decision by Meta to leave up a Facebook post that attacked an identifiable woman and compared her to a 'truck'. After the FOB brought the appeal to Meta's attention, the company reversed its original decision and removed the post: 'Dehumanizing Speech against a Woman', *Oversight Board* (Web Page, 27 June 2023) <https://oversightboard.com/decision/FB-VJ6FO5UY/>.

main ways. It is speech that is amplified on platforms, as an aspect of platforms' instrumental power or ability to influence communication through control of social media. It is also speech that is especially accommodated, and thus authorised, on platforms, as an aspect of platforms' structural power as significant channels for self-presentation and communication.

Platforms' corporate and profit imperatives underly this platforming of sex-based vilification and impede the effective regulation of such speech. Significantly, platforms seek to maintain control or influence over the conditions for their own regulation and governance, through their discursive construction of themselves as the only or most legitimate arbiters of the treatment of the speech they host. Related to this is a privileging of self-regulatory action in current laws and law reform proposals for platform governance, which, as we showed with reference to Facebook and the FOB, is both inadequate and inappropriate to mitigate the harms to women of platformed sex-based vilification. Moreover, self-regulation obfuscates the harms of platformed sex-based vilification by appearing to provide an internal governance solution that is not a solution and undermines potential for rendering platform power accountable for such speech. In doing so, it reinforces failures of existing laws to address such speech.

We argued that through the use of their discursive power to privilege self-regulation, platforms are thus responsible for discursively 'rubberstamping' ineffective or anti-feminist content moderation processes and outcomes in ways that re-enact platformed sex-based vilification's harms to women, as well as platforms' complicity in those harms. In this way, platformed sex-based vilification is also auspiced by platforms, as an aspect of their discursive power. We argued that this auspicing represents an additional layer of contempt for women, for which platforms currently are not but should be held accountable.

In lieu of a privileging of platform self-regulation, effectively addressing the harms of platformed sex-based vilification requires a multifaceted 'ecosystem' of legal and regulatory mechanisms. Such an ecosystem would constrain platform power and hold platforms responsible structurally with respect to the role their instrumental and infrastructural power play in the amplification, accommodation and authorisation of sex-based vilification and its harms to women. It would also hold platforms accountable for the role their discursive power plays in the auspicing of platformed sex-based vilification. Importantly, it would ensure that any self-regulatory responses to platformed sex-based vilification are sufficiently buttressed by laws providing real and systemic recourse to women and counteracting attempts by platforms to obfuscate or escape liability for their complicity in the harms of such speech. This is especially important given that some platforms appear to be abandoning significant aspects of even the minimal self-regulation they have previously subjected themselves to in this regard, such that the relevant harms will be further exacerbated,[274] with more platforms predicted to follow.[275] We leave these points for future work.

274 See above nn 36–8 in relation to Facebook and Meta.
275 See, eg, 'Meta's Misinformation Shift' (n 38).