Volume **7** (2) 2025 https://doi.org/10.5204/lthj.3848

A Misogynistic Glitch? A Feminist Critique of Algorithmic Content Moderation

Valentina Golunova
Maastricht University, The Netherlands

Abstract

In recent years, all leading social media platforms have integrated artificial intelligence (AI) into their content moderation workflows. An increasingly prevalent narrative suggests that algorithms capable of detecting and removing prohibited or harmful content are crucial for ensuring that marginalised groups have equal opportunities to participate in civic discourse. Drawing on feminist theory, this article disproves this narrative. It begins by unpacking the evolution of perceptions of algorithmic content moderation, shining a light on the recent departure from simplistic efficiency considerations towards a more compelling portrayal of AI as a means of fostering a more inclusive dialogue online. It then examines how the use of AI for detecting and removing violative material further marginalises – rather than empowers – women who are seeking to engage on social media. Apart from being unable to properly address online gender-based violence and misogyny, algorithms employed in content moderation often erroneously restrict women's lawful counter-speech, thus preventing them from contributing to public debate. The article concludes with brief reflections on how the inevitable expansion of technological solutions in content moderation could be aligned with feminist ideals.

***Keywords**:* Artificial intelligence (AI); content moderation; gender; algorithm; social media; misogyny.

1. Introduction

Content moderation is a crucial activity that allows social media platforms to counter abuse and promote a healthy civic discourse on their communication networks.[1] Over the past decade, platforms have increasingly relied on algorithms to make their content moderation processes more expeditious and efficient. Algorithms, which are typically powered by artificial intelligence (AI), are deployed to classify text, images or multimedia files in accordance with the set guidelines. Upon detecting an item that appears to violate these guidelines, an algorithm can immediately apply a relevant restriction, preventing or minimising its further dissemination. The advent of generative AI has stirred a new wave of fascination with technology's potential to address hate and abuse in the digital domain. Large Language Models (LLMs) are confidently presented as robust tools for interpreting and applying complex online speech rules.[2]

The sweeping advance of AI as one of the most prominent frontier technologies has given rise to many diverse narratives.[3] Remarkably, the advent of algorithmic content moderation has been accompanied by predominantly techno-optimistic accounts. There is a rapidly emerging contention that the deployment of AI in content moderation can promote more inclusive communication spaces by combatting abuse and toxicity, and elevating frequently silenced voices.[4] Yet a growing body of critical legal scholarship paints a much grimmer picture of this phenomenon. Much of this scholarship argues that an AI-driven

[1] Gillespie, Custodians of the Internet, 5–6.
[2] Weng, "Using GPT-4 for Content Moderation."
[3] Sartori, "A Sociotechnical Perspective."
[4] See, for example, Rieder, "The Fabrics of Machine Moderation," 4; Oh, "Does Algorithmic Content Moderation?" 2.

approach to addressing offensive or harmful content on social media encodes and perpetuates existing forms of oppression and discrimination. Some authors analyse algorithmic content moderation through a decolonial lens, claiming that it propagates rather than counters systemic racism.[5] Others have exposed the impact of AI in content moderation on vulnerable or marginalised groups. For instance, it has been shown that algorithms are insensitive towards distinct communicative practices or provocative visual representations of the LGBTQI+ community, fuelling a sense of exclusion.[6] Additionally, studies have found that AI-driven moderation practices discriminate against individuals of Palestinian descent based on specific location tags, catchphrases and images commonly used by them.[7] There is also a burgeoning line of literature addressing the impact of algorithmic content moderation on women. Some writings challenged the technology's potential to combat online gender-based violence,[8] whereas others highlighted how AI used to facilitate moderation tasks can further undercut female voices on social media platforms.[9] Yet a fully-fledged feminist critique of algorithmic content moderation, along with a coherent understanding of how a feminist perspective could be used to highlight its shortcomings, is still lacking.

This article critically interrogates the current narrative around algorithmic content moderation through the lens of feminist theory. Section 2 traces the evolution of narratives around algorithmic content moderation. It highlights how the initial portrayal of AI as a tool for lowering costs and increasing the efficiency of moderation has gradually been replaced by more ambitious claims that such tools can enhance the diversity and equality of public discourse online. Section 3 disconfirms this narrative by demonstrating that the increasing reliance on algorithmic content moderation is ill-suited for protecting and empowering women seeking to engage in public discourse online. It claims that algorithms are inadequate for detecting and removing content that harms women. Furthermore, by wrongfully restricting women's legitimate content, algorithms fail to meaningfully safeguard and promote their contribution to public discourse on social media. Despite these acute issues, section 4 grapples with the inevitability of the 'algorithmisation' of content moderation and examines several alternative options for instrumentalising technology to safeguard women's rights online. Section 5 concludes by casting a forward-looking perspective on the use of AI in content moderation and the possibilities for aligning it with the values of gender diversity and inclusivity in the digital environment.

This article makes three contributions. It furthers the criticism of the deeply entrenched postulation that technology has a democratising effect on society.[10] Additionally, it proposes a more complex theoretical lens through which algorithmic content moderation and its impact on democratic discourse should be studied. Finally, it provides tentative reflections on how technology could be reformed to end oppression and foster social justice. This article therefore responds to the call to operationalise the feminist literature to rethink the way AI could facilitate content moderation.[11]

2. The Narratives Around Algorithmic Content Moderation: From Efficiency to Inclusivity

Content moderation is notoriously tedious. The staggering growth in the number of social media users, escalating political polarisation, the proliferation of malign actors and other factors all contribute to a slew of undesirable content that is extremely difficult to eradicate or manage. Against this backdrop, automation is seen as one of the few viable alternatives for tackling illegal or harmful material on social media. In recent years, an AI-driven approach to content moderation has gained outstanding prominence: most decisions currently made by major social media platforms are either fully or partially automated.[12]

Law plays an instrumental role in incentivising online platforms to use AI to address problematic content made available by their users.[13] On the one hand, social media platforms are either implicitly or explicitly encouraged to use AI to tackle illegal material. For example, several pieces of EU legislation, such as the Directive on Copyright in the Digital Single Market and the Terrorist Content Regulation, contain provisions that require platforms to take *ex ante* measures against copyright-infringing content and content inciting or glorifying terrorist activities respectively, which necessitates the deployment of AI for more rapid and comprehensive screening of uploaded material.[14] The recent legislative reforms in countries such as Australia and

[5] Siapera, "AI Content Moderation"; Shahid, "Decolonizing Content Moderation."
[6] Dias Oliva, "Fighting Hate Speech"; Griffin, "The Heteronormative Male Gaze."
[7] Abokhodair, "Opaque Algorithms, Transparent Biases."
[8] Oh, "Does Algorithmic Content Moderation?"
[9] Gerrard, "Social Media Content Moderation"; Riccio, "A Techno-Feminist Perspective."
[10] Barbrook, "The Californian Ideology"; Laufer, "Algorithmic Displacement of Social Trust."
[11] Gerrard, "Social Media Content Moderation," 749.
[12] Kaushal, "Automated Transparency," 6–7.
[13] Golunova, "Silenced by Default," 3–7.
[14] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market [2019] OJ L 130/92, art 17(4); Regulation (EU) 2021/784 of 29 April 2021 on addressing the dissemination of terrorist content online [2021] OJ L 172/79, art 5. For a more detailed discussion see Castets-Renard, "Algorithmic Content Moderation."

India also point to increasing recognition of AI as a means of tackling illegal activity in the digital environment.[15] On the other hand, most social media laws around the world do not impose any accountability on social media platforms for the wrongful removal of legitimate content. For instance, Section 230 of the US Communications Decency Act provides exceptionally broad immunity from liability for interactive computer services when restricting user-generated content deemed 'obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected'.[16] Notably, the Digital Services Act – the most recent piece of EU legislation that revised the horizontal rules governing online intermediaries – requires online platforms to have due regard to fundamental rights when enforcing their terms and conditions.[17] In practice, however, social media platforms do not carry significant legal risks when their AI-driven moderation systems cause large-scale removal of content protected by freedom of expression.[18]

Given the crucial role of algorithmic content moderation in the context of platform governance and regulation, it has come to be surrounded by a diverse set of narratives. According to a widely made claim, the use of AI significantly increases the scalability, consistency and efficiency of moderation activities.[19] Algorithms are also believed to improve working conditions for human moderators, who are thereby relieved from the burden of reviewing some of the most heinous material.[20] Moreover, there is a common assertion that algorithmic content moderation benefits law enforcement by enabling prompter detection and investigation of criminal activity online.[21]

The advent of algorithmic content moderation and the regulators' rush to endorse it as a solution to the spread of illegal content online have inspired vigorous academic exploration, with some of the contributions specifically focusing on dismantling the misleading or overblown narratives around it. Katzenbach has drawn a parallel between the rush to implement algorithms in content moderation and the notions of the 'technological fix' and 'solutionism', coined by Volti and Morozov respectively, to describe blunt trust in technological affordances for addressing complex societal issues.[22] When analysing whether AI truly benefits workers, Griffin has pointed out that the continuous training of algorithms is dependent on extensive and poorly paid human labour typically sourced from the Global South, thus causing a heavier workload and worse working conditions.[23] One should also be conscious of the fact that the strong commitment to the use of AI in content moderation is tightly linked to the ongoing crisis of the Trust & Safety industry caused by large-scale layoffs in the tech sector. After reducing the number of qualified personnel in charge of reviewing moderation decisions, social media platforms claimed that algorithms could easily take over this function.[24] However, the recourse to AI merely obfuscates the platforms' failure to attract and retain talent that could develop and implement more comprehensive solutions for maintaining the safety and security of their services. The reliance on algorithms as instruments of public policy is equally flawed, as it could result in abuses of power and obstruct the foreseeability and fairness of restrictions on users' expression.[25]

Given the substantial criticism around the 'algorithmisation' of content moderation, social media platforms committed to shifting the emphasis from cost-effectiveness as the key advantage of algorithmic content moderation to its alleged societal benefits. According to one of the most recently emerged contentions, AI does not just help maintain the safety and security of platforms' communication networks by weeding out problematic content but also materially improves the quality of online conversations by uplifting commonly silenced groups. Meta, for instance, indicates that its innovative method Few-Shot Learner (FSL) is less dependent on a high load of training data and thereby ensures a more sensitive approach to moderation – especially when it comes to Arabic content, which is typically subject to erroneous restrictions.[26] Moreover, Spill – the social media platform founded in the aftermath of Elon Musk's purchase of Twitter (now X) – claims that its content moderation tool is powered by an LLM trained by Black, LGBTQI+, and other marginalised communities with the view of ensuring the accuracy

[15] Douek, "Australia's 'Abhorrent Violent Material' Law"; Gupta, "Evolving Scope of Intermediary Liability."
[16] 47 U.S. Code § 230.
[17] 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, art 14(4). For a more in-depth discussion of this provision, see Quintais, "Using Terms and Conditions."
[18] Krönke, "Artificial Intelligence and Social Media," 163.
[19] Gorwa, "Algorithmic Content Moderation," 2–5.
[20] Darbinyan, "The Growing Role of AI."
[21] Bloch-Wehba, "Content Moderation as Surveillance," 1313–1314.
[22] Katzenbach, "'AI Will Fix This'," 2.
[23] Griffin, "Algorithmic Content Moderation."
[24] Malik, "ByteDance Lays Off Hundreds."
[25] Castets-Renard, "Algorithmic Content Moderation," 311–313.
[26] Meta AI, "Harmful Content Can Evolve Quickly."

and fairness of its outcomes.[27] Accordingly, algorithmic content moderation is increasingly presented as a means of achieving a more robust and inclusive public debate.

The seemingly irreversible expansion of automation in content moderation raises the question of whether technology does, or could potentially, protect and elevate the voices of women as one of the vulnerable groups. Notably, many efforts to build and put in place AI-driven solutions for content moderation grew out of a desire to mitigate online gender-based violence and misogyny – some of the most urgent and delicate harms arising on social media. Google/Jigsaw's Perspective application programming interface (API) – a pioneering AI-driven system for detecting toxic language – was developed in response to the Gamergate campaign that targeted women and gender-diverse people in the video game industry.[28] At various points in time, online platforms have also committed to technological solutions to combat a wide range of gender-related threats, such as troll attacks and non-consensual intimate deepfakes (NCIDs).[29] At the same time, oppression and silencing of women on social media persist. The 2024 report of the UN Secretary-General revealed that between 16 and 58 per cent of women and girls had experienced technology-facilitated violence, with certain demographics (based on age and country of residence) being particularly susceptible to it.[30] This calls for a critical analysis of an AI-driven approach to content moderation.

3. Constructing a Feminist Lens on Algorithmic Content Moderation

The critical literature on the relationship between technology and women aspires to uncover how the dynamics of gender power relations manifest through scientific progress. The postmodernist work of Haraway set the stage for interrogating the role of sex, gender and identity against the backdrop of the increasing entanglement of human bodies with artificial objects.[31] Haraway's notion of 'cyborg' denotes the opportunity to break away from the traditional notions of masculinity and femininity. While this provocative vision attracted criticism, including from other feminist scholars,[32] it sparked a rich debate on the role of digitalisation as an economic and social phenomenon in gender discourse.[33]

Drawing on this literature, this section develops a theoretical framework for analysing and critiquing the alleged potential of algorithmic content moderation to ensure fairer and more inclusive communication spaces for women. It acknowledges the complexity of this phenomenon by framing it as both an essential form of social media governance as well as a process enabled by AI-driven tools as technological artefacts. It is the intersection of these two facets and the compound effects they produce that truly illuminates the implications of the use of AI in content moderation for women's expression online. To properly capture this intersection, this section synthesises two lines of feminist scholarship focusing on platform governance and AI respectively.

3.1 Algorithms as Platform Governance Mechanisms

Algorithms are deeply embedded in the architecture of social media platforms. By effectuating platforms' policies, algorithms are one of the key instruments of platform governance – a complex blend of structures and processes shaping the interactions between actors within the platforms' ecosystem.[34] A feminist perspective on algorithmic content moderation can therefore be informed by critical studies focusing on various facets of platform governance. Despite the initial hope that the advent of social media will ensure wider participation of marginalised communities, including women, in civic discourse and cultural production, feminist scholars have continuously exposed how its features contribute to sustaining social hierarchies and subordination. Some writings have investigated how social media platforms reinforce gender normativity.[35] The work of Gerrard and Thornham, for instance, exposes 'sexist assemblages', or various interlinked components of platform governance, including keyword and hashtag searches, platforms' community guidelines and recommender systems that espouse gender bias and inequality.[36]

[27] Hendrix, "Is Generative AI the Answer?"
[28] Rieder, "The Fabrics of Machine Moderation," 4.
[29] See, among others, Sinders, "Technology Alone"; Bickert, "Our Approach to Labeling."
[30] UN Secretary-General, Intensification of Efforts.
[31] Haraway, "A Manifesto for Cyborgs."
[32] See, for instance, Wajcman, Technofeminism, 101 (highlighting the limited utility of Haraway's work for 'practical emancipatory politics'); Gillis, "Feminist Criticism and Technologies of the Body," 329 (drawing attention to the neglect of race in the concept of 'cyborg').
[33] Van Loon, "Technological Sensibilities and the Cyberpolitics of Gender"; Sandoval, "New Sciences"; Suchman, "Feminist STS and the Sciences of the Artificial."
[34] Gorwa, "What is Platform Governance?" 856–857.
[35] See, among others, Carstensen, "Gender and Social Media"; Bivens, "Baking Gender."
[36] Gerrard, "Content Moderation."

Other contributions have looked at the role of platform governance mechanisms in fuelling or normalising misogyny. Specifically, they emphasise the platforms' persistent failure to address abusive language against women,[37] curb the expansion of the 'manosphere'[38] or mitigate the weaponisation of new technologies against women.[39] The authors typically attribute these failures to the platforms' insensitivity towards deeply entrenched gender power struggles and vulnerabilities.

Lastly, another important strand of literature has investigated the opportunities and barriers faced by women in expressing themselves on social media. Nurik, for instance, introduced the term 'gender-based censorship' to describe the segmentation of users depending on their perceived benefit to platforms.[40] In challenging the definition of social media as the new public square within the meaning of Habermas's theory, Franks succinctly points out that social media platforms are privately owned, profit-oriented entities, which leads them to seek to maximise engagement rather than open and inclusive dialogue on societally important issues.[41] Building on the works of Fraser and Crenshaw, Galpin further unpacks the asymmetries of social media engagement, whereby women – especially those with intersectional identities – continue to face barriers when contributing to political discourse.[42] Accordingly, the design and functioning of platform governance are often argued to be at odds with the key purpose of social media: to liberalise communication and unleash creativity on a global scale.

3.2 Algorithms as Technological Artefacts

Critical platform governance studies present an already complex set of insights that can be leveraged for analysing algorithmic content moderation. However, by relying solely on these studies – no matter how insightful they may be – one cannot envision a complete picture of how AI may or may not protect and empower women on social media platforms. By seeing algorithms only as part of the platforms' infrastructure, one risks overlooking other important stages of their life-cycle, such as development, training and iteration.[43] Hence, framing algorithms as technological artefacts is equally crucial for investigating their impact on women's participation in public discourse online.

Scholarly writings across various fields, such as sociology, philosophy and political science, depart from the premise that technological artefacts are not neutral, but rather reflect and reproduce gender patterns and relations within broader society. Wajcman, for instance, masterfully bridged feminist theories and science and technology studies (STS) to expose the male domination and bias in technological development.[44] Coming from a history of science perspective, Adam has applied feminist epistemology to critique the way AI obstructs women's knowledge production.[45] These and other related early academic works have laid the foundation for modern-day critiques of AI, which is widely argued to propagate patriarchy and masculinity.[46] This argument is supported by the reference to the persistent exclusion of women from participating in the designing, testing and implementation of AI.[47] As AI innovation continues to be concentrated in men's hands, the perspectives of marginalised groups, including women, remain neglected.

The inequalities that manifest themselves in the process of AI development inevitably influence the quality and fairness of outputs produced by relevant tools.[48] Women, particularly those with marginalised backgrounds, continue to be under-represented in training datasets. Cutting-edge research by Buolamwini and Gebru showed, for example, that, due to inadequate dataset composition, commercial facial analysis algorithms performed the worst on darker-skinned women.[49] Gender biases and stereotypes also creep into the AI training process. Even where more advanced algorithms are allegedly less dependent on manual human labelling and are capable of learning independently from a large corpus of data, the ubiquity of prejudices in official resources and the media continue to shape their outputs in ways that perpetuate the marginalisation of certain groups.[50] The domination of a few powerful corporations standing at the forefront of the AI revolution and their ignorance of the issues of gender segregation results in a lack of incentive to fight the injustices 'baked' into technological artefacts. In this regard,

[37] Barker, "Online Misogyny."
[38] Trott, "Operationalising 'Toxicity'."
[39] Kira, "Deepfakes."
[40] Nurik, "'Men are Scum,'" 2893.
[41] Franks, "Beyond the Public Square," 428–429.
[42] Galpin, "At the Digital Margins?" 164–166.
[43] Khan, "Framing Online Speech Governance," 39.
[44] Wajcman, Feminism Confronts Technology.
[45] Adam, Artificial Knowing.
[46] See, for example, Wajcman, "Feminism Confronts AI," 52.
[47] Young, "Mind the Gender Gap."
[48] Manasi, "Mirroring the Bias"; Broussard, More than a Glitch.
[49] Buolamwini, "Gender Shades."
[50] Ananya, "AI Image Generators," 723.

Noble coins the term 'algorithmic oppression' to describe the offensive and discriminatory results generated by search algorithms and their role in perpetuating racism and sexism.[51] Therefore, the existing critiques of AI as a technological artefact offer an additional valuable standpoint on the role of algorithmic content moderation in ensuring a healthy and diverse public debate online.

4. Challenging the Promise of Algorithmic Content Moderation to Protect and Empower Women on Social Media

The introduction of AI in the content moderation pipeline sparks intricate sociotechnical effects. Such effects are determined by the complex nature of algorithms employed for moderation purposes, which are an integral part of overarching platform governance mechanisms as well as independent technological artefacts. By blending the insights from the two strands of literature addressed above, this section embraces this complexity to unravel AI's capacity to secure and amplify women's participation in civic discourse on social media. It examines three key algorithmically assisted interventions explicitly purported to address content or behaviour aimed at harming, discrediting or silencing women: the automated removal of prohibited material; the detection of coordinated inauthentic campaigns; and the regulation of the visibility of harmful content that does not explicitly violate the applicable rules. Additionally, the overwhelming reliance on AI in content moderation often prompts erroneous restrictions on women's posts, in which they expose or resist abuse, which completely defies its purpose. The devastating impact of algorithms on women's expression refutes the growing narrative around their potential for ensuring a more open and diverse public discourse on social media.

4.1 Algorithmic Removal of Prohibited Content

One of the most common applications of AI in content moderation involves the identification and removal of material that is either illegal or prohibited by platforms' own terms and conditions.[52] Its main advantage lies in its ability to support a proactive response towards user violations. Unlike human moderators, who struggle to keep up with the ever-increasing amount of content hosted by platforms, algorithms can weed out inappropriate items at scale.

At first glance, the AI-assisted search for prohibited content appears to be beneficial for tackling prohibited content targeting women. Research shows that mechanisms of individual redress (such as the possibility to flag content as violating terms and conditions) often prove futile as they shift the burden from platforms, whose infrastructure enables abuse and harassment, to users, who are often reluctant to resort to these mechanisms due to a lack of trust in their effectiveness or fairness.[53] Algorithms are alleged to compensate for this issue by rapidly removing relevant material before it is widely viewed or disseminated, thereby causing significant harm to women's well-being or reputation.

However, the use of AI for eliminating inappropriate content on social media raises multiple concerns. Above all, it does not offer a reliable solution for combatting online gender-based violence. The accuracy of algorithms is directly dependent on the quality and precision of label categories used to classify content. However, platforms are notorious for maintaining an unsophisticated and reactionary approach to defining prohibited content.[54] While most of them emphasise their commitment to gender equality and inclusivity, their terms and conditions do not capture the important nuances of qualifying speech as misogynistic.[55] Thus, offensive content is often claimed to merely constitute 'humour'[56] or 'parody'.[57] Furthermore, platforms regularly fail to promptly update their policies to address novel types of harms, such as NCID.[58] The flaws of platforms' terms and conditions are exacerbated by the technical limitations of AI tools that cannot analyse the intention behind the content or the context around it. For instance, algorithms are largely powerless when users conceal hateful or abusive content by using 'algospeak' – special coded language that allows to evade automated restrictions.[59] As a result, much abusive material ends up undetected by AI.

The combination of inadequate platforms' policies and the context- and nuance-blind nature of algorithms used to enforce them not only undercuts the effectiveness of an AI-driven approach to countering inappropriate content but also results in the

[51] Noble, Algorithms of Oppression.
[52] Gonçalves, "Algorithmic Moderation," 530.
[53] See, for instance, Vaccaro, "At the End of the Day," 16; Eder, "Making Systemic Risk Assessments Work," 7.
[54] See, among others, Suzor, "Human Rights by Design," 94; Nourooz Pour, "Voices and Values," 13.
[55] Richardson-Self, "Woman-Hating" (highlighting the distinction between sexist and intradivisional misogynistic speech).
[56] Chemaly, "Facebook's Big Misogyny Problem."
[57] Milmo, "Elon Musk Accused of Spreading 'Lies'."
[58] Kira, "Deepfakes."
[59] Levine, "These TikTok Accounts."

wrongful restriction of legitimate content uploaded by women. In addition to their well-documented intolerance of images of breastfeeding and menstruation,[60] platforms systematically remove content aimed at raising awareness about gender-based violence and exposing the role of men in perpetuating it.[61] Following the initial controversy around the moderation of the phrase 'men are scum' in 2017, the issue has resurfaced in a recent case before Meta's Oversight Board – an independent body tasked with reviewing content moderation decisions on Facebook, Instagram and Threads.[62] The Oversight Board reversed Meta's decision to remove posts denouncing violence against women that were flagged by an algorithm as violating its hate speech policy. This example reveals how platforms' long-standing ignorance of the issues around women's expression can seep into the design of their algorithms and be further exacerbated by their limited ability to recognise the true meaning and intent behind specific posts.

4.2 Algorithmic 'Deplatforming'

AI-driven solutions are employed not only for the identification of prohibited content but also for the detection and termination of problematic accounts. The latter measure (also known as 'deplatforming') is applied whenever a certain account persistently violates the applicable rules, engages in an inauthentic behaviour, or belongs to a list of dangerous organisations or individuals. At first glance, the purported benefit of AI is, once again, obvious. Algorithms can swiftly analyse the accounts' entire activity and disable them before they cause significant harm. Accordingly, automated 'deplatforming' is widely used to address misogynistic troll campaigns as well as restrict popular accounts spreading hateful sexist content.

Yet research shows that the use of AI for large-scale termination of accounts has limited effectiveness in cultivating more favourable conditions for public debate on social media. More notably, 'deplatforming' does not fully eliminate problematic behaviour but simply makes users migrate to different platforms.[63] For example, platforms such as YouTube and TikTok have pledged to use algorithms to terminate accounts replicating the profile of Andrew Tate – the social media influencer permanently banned for his misogynistic views.[64] At the same time, his persona and rhetoric continue to thrive on other platforms (including X, where his account was reinstated by Elon Musk shortly after he bought the platform in 2022).

Furthermore, platforms have been reported to apply automated 'deplatforming' in a highly unfair and disparate manner. In her fascinating autoethnography, Are details her experience with her pole dancing instructor's accounts being automatically disabled on Instagram and TikTok.[65] In developing the concept of 'automated powerlessness', she notes the lack of transparency around the reasons for 'deplatforming', safeguards against malicious flagging and the necessary affordances for recovering access to the account.[66] The issue is aggravated by the pervasive opacity of algorithms. It is often nearly impossible to determine why they produce certain outputs. For instance, the Facebook account of Australian feminist activist Clementine Ford has been suspended multiple times for exposing sexist threats received from other users or pushing back against troll attacks.[67] At the same time, the original content targeting Ford was not subject to automated removal. Therefore, AI-assisted 'deplatforming' does not just offer a merely piecemeal solution to online gender-based violence, but also affects women's legitimate counter-speech.

4.3 Algorithmic Demotion of Undesirable Content

For a long time, the removal of content and accounts violating the law or platforms' policies has been a key means of content moderation. Increasingly, however, platforms are more concerned with the visibility of, rather than access to, content.[68] When certain content is *demoted*, it remains available online but is downgraded in users' feeds and not promoted to a new audience through the recommender systems. This measure is usually applied to material that does not violate the applicable rules but is nevertheless harmful or disturbing.

AI-driven content demotion is increasingly seen as a strategy for addressing the amplification of content that does not meet the threshold of incitement to gender-based violence but could nevertheless be offensive to women.[69] Nonetheless, this strategy frequently clashes with platforms' desire to serve personalised, engagement-prone content in order to keep users drawn to their

[60] Faust, "Hair, Blood and the Nipple."
[61] Marshall, "Algorithmic Misogynoir," 7.
[62] Oversight Board, "Violence Against Women."
[63] Cima, "The Great Ban."
[64] Sung, "Andrew Tate Banned."
[65] Are, "An Autoethnography."
[66] Are, "An Autoethnography," 836.
[67] Sims, "This Woman is Highlighting."
[68] Gillespie, "Do Not Recommend?," 1–2.
[69] See, for instance, Howard, "Remove or Reduce," 14–15 (exploring the demotion of 'borderline' hate speech).

services, which is a crucial driver of their advertising revenues. In this context, the women's and girls' rights advocate Lucina Di Meco points to the phenomenon of 'attention economy' as the primary reason for the platforms' failure to combat online gendered disinformation.[70] Sexist prejudices 'baked' into AI's design further undermine the efforts to limit the visibility of undesirable content. Thus, in a recent experiment by *Guardian Australia*, the AI-driven recommender systems of Facebook and Instagram prioritised misogynistic content on the explore pages of dummy male accounts that were linked to unused credentials on empty devices.[71] Such an outcome can be attributed to the alarming assumption that hateful or discriminatory content targeting women is most appealing to men within certain demographics.

In a similar vein to AI-driven content removal and 'deplatforming', the demotion of undesirable content impacts women's expression on social media. As succinctly argued by Riemer and Peter, it gives rise to the notion of 'algorithmic audiencing', whereby the selection of users who see specific content is determined in arbitrary and opaque ways, which risk leaving certain messages unheard.[72] There is a widely documented practice of 'shadowbanning' – the opaque restriction of the content's reach – of women's posts on politically important topics (such as abortions)[73] or images of female bodies.[74] Accordingly, the amalgamation of platforms' attention-driven business model with gender bias engrained in the algorithms that are deployed seriously curtails their ambition to facilitate more diverse and inclusive spaces for public debate online.

5. Concluding Remarks

As the use of AI in content moderation expands dramatically, it faces increasing pushback. As a result, the narratives accompanying it are evolving and becoming more elaborate. This article has put a spotlight on the gradual transition from a rather short-sighted promise of cost-effectiveness towards a more compelling claim that algorithmic content moderation can make platforms more inclusive spaces for democratic discourse. At the same time, it has been shown why recent technological advancements fail to enable women to engage on social media without becoming a target of hateful or derogatory content or facing unjustified restrictions on their expression. Algorithms employed in content moderation generate effects stemming from both their integration into the overarching system of platform governance and from the flaws of the underlying technology. Such effects are intertwined and mutually enhancing, as their interplay reinforces exclusion and discrimination. The growing assertion that recent technological achievements in the field of AI can help to create more equal opportunities to contribute to the public debate on social media is, therefore, misguided.

Despite taking a critical stance on algorithmic content moderation, this article does not aim to fully discredit the technology's potential for elevating women's voices on social media. For the foreseeable future, the AI-driven interventions discussed in this article will remain necessary for regulating the information flow and ensuring user safety. At the same time, liberating AI from the confines of platforms' infrastructure, such as by offering customisable automated tools allowing individuals to remove or hide abuse and harassment from their feeds, could help to unlock its potential for safeguarding the quality and diversity of public discourse online. The anti-harassment tool Block Party, developed by American software engineer Tracy Chou, is a notable example of how technology could be leveraged to enhance user agency without reinforcing platform power.[75] In addition, the mitigation of deeply entrenched opacity and bias demands a structural reform of the AI industry and further research into the strategies for curating training datasets and de-biasing techniques. For instance, cutting-edge studies on encoding societal values into AI offer promising approaches to ensuring equality and protecting human rights by design.[76]

Admittedly, the success of both these initiatives could be seriously jeopardised by the current right-wing turn of tech companies, spurred by the second Trump administration in the United States.[77] The shift towards more lax moderation policies and the discontinuation of diversity, equity and inclusion (DEI) programs underscore that social media platforms are more reluctant than ever to stand for the interests of vulnerable groups. However, advocating for meaningful change to the architecture and functioning of content moderation is paramount, as simply delegating the vital task of creating more equitable online spaces to AI obfuscates the fundamental problems inherent to platform governance and technology production. A critical interrogation of the narrative concerning AI's ability to promote healthy and inclusive conversations on social media serves as an essential first step to thinking deeply about the alternative means of furthering this vital cause.

[70] Di Meco, "Monetizing Misogyny," 22.
[71] Taylor, "We Unleashed Facebook."
[72] Riemer, "Algorithmic Audiencing."
[73] Seitz, "Instagram Hides Some Posts."
[74] Mauro, "'There is No Standard.'"
[75] See, for instance, Bond, "Block Party Aims To Be A 'Spam Folder' For Social Media Harassment."
[76] Bernstein, Christin, and Hancock, "Tuning Our Algorithmic Amplifiers."
[77] Kaplan, "More Speech and Fewer Mistakes."

Acknowledgements
The author wishes to thank Dr Henrique Marcos and Syamsuriatina Ishak, the organisers of the conference *Narratives, Frontier Technologies & The Law* (Maastricht University, 30 October–1 November 2024), who created a unique and stimulating space for presenting the early version of this article. Sincere gratitude also goes to the participants of this conference for inspiring discussions. The author would also like to thank the anonymous reviewers for their helpful comments and suggestions.

Bibliography

Primary sources

47 U.S. Code § 230.

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market [2019] OJ L 130/92.

Regulation (EU) 2021/784 of 29 April 2021 on addressing the dissemination of terrorist content online [2021] OJ L 172/79.

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.

United Nations Secretary-General. *Intensification of Efforts to Eliminate All Forms of Violence Against Women and Girls: Technology-Facilitated Violence against Women and Girls*. A/79/500. New York: United Nations, 2024. https://www.unwomen.org/sites/default/files/2024-10/a-79-500-sg-report-ending-violence-against-women-and-girls-2024-en.pdf.

Secondary sources

Abokhodair, Norah, Yarden Skop, Sarah Rüller, Konstantin Aal and Houda Elmimouni. "Opaque Algorithms, Transparent Biases: Automated Content Moderation During the Sheikh Jarrah Crisis." *First Monday* 29, no 4 (2024). https://doi.org/10.5210/fm.v29i4.13620.

Adam, Alison. *Artificial Knowing: Gender and the Thinking Machine*. London: Routledge, 1998.

Ananya. "AI Image Generators Often Give Racist and Sexist Results: Can They Be Fixed?" *Nature* 627, no 8005 (2024): 722–725. https://doi.org/10.1038/d41586-024-00674-9.

Are, Carolina. "An Autoethnography of Automated Powerlessness: Lacking Platform Affordances in Instagram and TikTok Account Deletions." *Media, Culture & Society* 45, no 4 (2023): 822–840. https://doi.org/10.1177/01634437221140531.

Barbrook, Richard and Andy Cameron. "The Californian Ideology." *Science as Culture* 6, no 1 (1996): 44–72. https://doi.org/10.1080/09505439609526455.

Barker, Kim and Olga Jurasz. "Online Misogyny: A Challenge for Digital Feminism?" *Journal of International Affairs* 72, no 2 (2019): 95–114.

Bernstein, Michael S., Angèle Christin and Jeffrey T. Hancock. "Tuning Our Algorithmic Amplifiers: Encoding Societal Values into Social Media AIs," October 20, 2023. https://hai.stanford.edu/news/tuning-our-algorithmic-amplifiers-encoding-societal-values-social-media-ais.

Bickert, Monika. "Our Approach to Labeling AI-Generated Content and Manipulated Media." Meta (blog), April 5, 2024. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media.

Bivens, Rena and Oliver L. Haimson. "Baking Gender into Social Media Design: How Platforms Shape Categories for Users and Advertisers." *Social Media + Society* 2, no 4 (2016). https://doi.org/10.1177/2056305116672486.

Bloch-Wehba, Hannah. "Content Moderation as Surveillance." *Berkeley Technology Law Journal* 36 (2022): 1297. https://doi.org/10.15779/Z389C6S202.

Bond, Shannon. "Block Party Aims to Be A 'Spam Folder' for Social Media Harassment." *NPR*, February 23, 2021. https://www.npr.org/2021/02/23/970300911/block-party-aims-to-be-a-spam-folder-for-social-media-harassment.

Broussard, Meredith. *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. Cambridge, MA: MIT Press, 2023.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. PMLR, 2018. https://proceedings.mlr.press/v81/buolamwini18a.html.

Carstensen, Tanja. "Gender and Social Media: Sexism, Empowerment, or the Irrelevance of Gender?" In *The Routledge Companion to Media & Gender*, edited by Cynthia Carter, Linda Steiner and Lisa Mclaughlin, 483–492. London: Routledge, 2014.

Castets-Renard, Céline. "Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement." *University of Illinois Journal of Law, Technology & Policy* 2, no 2 (2020): 283. https://doi.org/10.2139/ssrn.3535107.

Chemaly, Soraya. "Facebook's Big Misogyny Problem." *The Guardian*, April 18, 2013. https://www.theguardian.com/commentisfree/2013/apr/18/facebook-big-misogyny-problem.

Cima, Lorenzo, Amaury Trujillo, Marco Avvenuti and Stefano Cresci. "The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit." In *Companion Proceedings of the 16th ACM Web Science Conference*, 2024, 85–93. https://doi.org/10.1145/3630744.3663608.

Darbinyan, Rem. "The Growing Role of AI In Content Moderation." *Forbes*, June 14, 2022. https://www.forbes.com/sites/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation.

Dias Oliva, Thiago, Dennys Marcelo Antonialli and Alessandra Gomes. "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online." *Sexuality & Culture* 25, no 2 (2021): 700–732. https://doi.org/10.1007/s12119-020-09790-w.

Di Meco, Lucina. "Monetizing Misogyny: Gendered Disinformation and the Undermining of Women's Rights and Democracy Globally," February 2023. https://she-persisted.org/wp-content/uploads/2023/02/ShePersisted_MonetizingMisogyny.pdf.
Douek, Evelyn. "Australia's 'Abhorrent Violent Material' Law: Shouting 'Nerd Harder' and Drowning Out Speech." *Australian Law Journal* 94 (2020): 41.
Eder, Niklas. "Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation." SSRN Scholarly Paper, June 26, 2023. https://doi.org/10.2139/ssrn.4491365.
Faust, Gretchen. "Hair, Blood and the Nipple: Instagram Censorship and the Female Body." In *Digital Environments. Ethnographic Perspectives Across Global Online and Offline Spaces*. Transcript Verlag, 2017. https://www.genderopen.de/handle/25595/499.
Franks, Mary Anne. "Beyond the Public Square: Imagining Digital Democracy." *Yale Law Journal Forum* 131 (2021): 427–453.
Galpin, Charlotte. "At the Digital Margins? A Theoretical Examination of Social Media Engagement Using Intersectional Feminism." *Politics and Governance* 10, no 1 (2022): 161–171. https://doi.org/10.17645/pag.v10i1.4801.
Gerrard, Ysabel. "Social Media Content Moderation: Six Opportunities for Feminist Intervention." *Feminist Media Studies* 20, no 5 (2020): 748–751. https://doi.org/10.1080/14680777.2020.1783807.
Gerrard, Ysabel and Helen Thornham. "Content Moderation: Social Media's Sexist Assemblages." *New Media & Society* 22, no 7 (2020): 1266–1286. https://doi.org/10.1177/1461444820912540.
Gillespie, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press, 2018.
Gillespie, Tarleton. "Do Not Recommend? Reduction as a Form of Content Moderation." *Social Media + Society* 8, no 3 (2022): 1. https://doi.org/10.1177/20563051221117552.
Gillis, Stacy. "Feminist Criticism and Technologies of the Body." In *A History of Feminist Literary Criticism,* edited by Susan Sellers and Gill Plain, 322–335. Cambridge: Cambridge University Press, 2007. https://doi.org/10.1017/CBO9781139167314.
Golunova, Valentina. "Silenced by Default: Algorithmic Content Moderation and Freedom of Expression in the European Union." Maastricht University, 2024. https://doi.org/10.26481/dis.20240422vg.
Gonçalves, João, and Ina Weber. "Algorithmic Moderation: Contexts, Perceptions, and Misconceptions." In *Handbook of Critical Studies of Artificial Intelligence*, edited by Simon Lindgren, 528–37. Cheltenham: Edward Elgar, 2023. https://doi.org/10.4337/9781803928562.00055.
Gorwa, Robert. "What is Platform Governance?" *Information, Communication & Society* 22, no 6 (2019): 854–71. https://doi.org/10.1080/1369118X.2019.1573914.
Gorwa, Robert, Reuben Binns and Christian Katzenbach. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society*, January–June 2020: 1–15. https://doi.org/10.1177/2053951719897945.
Griffin, Rachel. "Algorithmic Content Moderation Brings New Opportunities and Risks" (blog). *Centre for International Governance Innovation*, October 23, 2023. https://www.cigionline.org/articles/algorithmic-content-moderation-brings-new-opportunities-and-risks.
Griffin, Rachel. "The Heteronormative Male Gaze: Experiences of Sexual Content Moderation Among Queer Instagram Users in Berlin." *International Journal of Communication* 18 (2024): 1266–1288. https://sciencespo.hal.science/hal-04617175v1.
Gupta, Indranath and Lakshmi Srinivasan. "Evolving Scope of Intermediary Liability in India." *International Review of Law, Computers & Technology* 37, no 3 (2023): 294–324. https://doi.org/10.1080/13600869.2022.2164838.
Haraway, Donna. "A Manifesto for Cyborgs: Science, Technology and Socialist Feminism in the 1980s." *Socialist Review* 80, no 2 (1985): 65–108.
Hendrix, Paul M. Barrett, Justin. "Is Generative AI the Answer for the Failures of Content Moderation?" *Just Security*, April 3, 2024. https://www.justsecurity.org/94118/is-generative-ai-the-answer-for-the-failures-of-content-moderation.
Howard, Jeffrey W., Beatriz Kira and Louisa Bartolo. "Remove or Reduce: Demotion, Content Moderation, and Human Rights." SSRN Scholarly Paper, July 11, 2024. https://doi.org/10.2139/ssrn.4891835.
Kaplan, Joel. "More Speech and Fewer Mistakes." Meta Newsroom, January 7, 2025. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes.
Katzenbach, Christian. "'AI Will Fix This': The Technical, Discursive, and Political Turn to AI in Governing Communication." *Big Data & Society* July-December 2021: 1. https://doi.org/10.1177/20539517211046182.
Kaushal, Rishabh, Jacob van de Kerkhof, Catalina Goanta, Gerasimos Spanakis and Adriana Iamnitchi. "Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database." arXiv, May 3, 2024. https://doi.org/10.48550/arXiv.2404.02894.
Khan, Mehtab. "Framing Online Speech Governance as an Algorithmic Accountability Issue." *Indiana Law Journal* 99 (2024): 37–61. https://www.repository.law.indiana.edu/ilj/vol99/iss5/3.

Kira, Beatriz. "Deepfakes, the Weaponisation of AI Against Women and Possible Solutions." *Verfassungsblog* (blog). June 3, 2024. https://verfassungsblog.de/deepfakes-ncid-ai-regulation.

Krönke, Christof. "Artificial Intelligence and Social Media." In *Regulating Artificial Intelligence*, edited by Thomas Wischmeyer and Timo Rademacher, 145–173. Cham: Springer, 2020.

Laufer, Benjamin and Helen Nissenbaum. "Algorithmic Displacement of Social Trust" (blog). Knight First Amendment Institute at Colombia University, November 29, 2023. http://knightcolumbia.org/content/algorithmic-displacement-of-social-trust.

Levine, Alexandra S. "These TikTok Accounts are Hiding Child Sexual Abuse Material in Plain Sight." *Forbes*, December 7, 2022. https://www.forbes.com/sites/alexandralevine/2022/11/11/tiktok-private-csam-child-sexual-abuse-material.

Malik, Aisha. "ByteDance Lays Off Hundreds of TikTok Employees in Shift to AI Content Moderation." *TechCrunch*, October 11, 2024. https://techcrunch.com/2024/10/11/bytedance-lays-off-hundreds-of-tiktok-employees-in-shift-to-ai-content-moderation.

Manasi, Ardra, Subadra Panchanadeswaran, Emily Sours and Seung Ju Lee. "Mirroring the Bias: Gender and Artificial Intelligence." *Gender, Technology and Development* 26, no 3 (December 1, 2022): 295–305. https://doi.org/10.1080/09718524.2022.2128254.

Marshall, Brandeis. "Algorithmic Misogynoir in Content Moderation Practice." Heinrich-Böll-Stiftung European Union and Heinrich-Böll-Stiftung Washington, DC, June 2021. https://eu.boell.org/sites/default/files/2021-06/HBS-e-paper-Algorithmic-Misogynoir-in-Content-Moderation-Practice-200621_FINAL.pdf.

Mauro, Gianluca and Hilke Schellmann. "'There is No Standard': Investigation Finds AI Algorithms Objectify Women's Bodies." *The Guardian*, February 8, 2023. https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies.

Meta AI. "Harmful Content Can Evolve Quickly. Our New AI System Adapts to Tackle It." August 12, 2021. https://ai.meta.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it.

Milmo, Dan. "Elon Musk Accused of Spreading 'Lies' over Doctored Kamala Harris Video." *The Guardian*, July 29, 2024. https://www.theguardian.com/technology/article/2024/jul/29/elon-musk-accused-of-spreading-lies-over-kamala-harris-video.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press, 2018.

Nourooz Pour, Hesam. "Voices and Values: The Challenging Odyssey of Meta to Harmonize Human Rights with Content Moderation." *International Journal of Law and Information Technology* 32, no 1 (2024). https://doi.org/10.1093/ijlit/eaae009.

Nurik, Chloe. "'Men are Scum': Self-Regulation, Hate Speech, and Gender-Based Censorship on Facebook." *International Journal of Communication* 13 (2019): 21.

Oh, Dayei and John Downey. "Does Algorithmic Content Moderation Promote Democratic Discourse? Radical Democratic Critique of Toxic Language AI." *Information, Communication & Society* (2024): 1–20. https://doi.org/10.1080/1369118X.2024.2346531.

Oversight Board. "Violence Against Women," July 12, 2023. https://www.oversightboard.com/decision/ig-h3138h6s.

Quintais, João Pedro, Naomi Appelman and Ronan Ó Fathaigh. "Using Terms and Conditions to Apply Fundamental Rights to Content Moderation." *German Law Journal* 24, no 5 (2023): 881. https://doi.org/10.2139/ssrn.4286147.

Riccio, Piera and Nuria Oliver. "A Techno-Feminist Perspective on the Algorithmic Censorship of Artistic Nudity." *Hertziana Studies in Art History* 3 (2024). https://doi.org/10.48431/hsah.0310.

Richardson-Self, Louise. "Woman-Hating: On Misogyny, Sexism, and Hate Speech." *Hypatia* 33, no 2 (2018): 256–272. https://doi.org/10.1111/hypa.12398.

Rieder, Bernhard and Yarden Skop. "The Fabrics of Machine Moderation: Studying the Technical, Normative, and Organizational Structure of Perspective API." *Big Data & Society* 8, no 2 (2021): 1. https://doi.org/10.1177/20539517211046181.

Riemer, Kai and Sandra Peter. "Algorithmic Audiencing: Why We Need to Rethink Free Speech on Social Media." *Journal of Information Technology* 36, no 4 (2021): 409–426. https://doi.org/10.1177/02683962211013358.

Sartori, Laura and Andreas Theodorou. "A Sociotechnical Perspective for the Future of AI: Narratives, Inequalities, and Human Control." *Ethics and Information Technology* 24, no 1 (2022): 4. https://doi.org/10.1007/s10676-022-09624-3.

Seitz, Amanda. "Instagram Hides Some Posts That Mention Abortion." *AP News*, June 29, 2022. https://apnews.com/article/technology-ac5da9efe2e200f26ff7702df1496e38.

Shahid, Farhana, and Aditya Vashistha. "Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?" In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. Hamburg: ACM, 2023. https://doi.org/10.1145/3544548.3581538.

Siapera, Eugenia. "AI Content Moderation, Racism and (de)Coloniality." *International Journal of Bullying Prevention* 4, no 1 (2022): 55–65. https://doi.org/10.1007/s42380-021-00105-7.

Sims, Alexandra. "This Woman is Highlighting Facebook's Ridiculous Double Standards." *The Independent*, March 29, 2016. https://www.independent.co.uk/tech/feminist-writer-clementine-ford-is-highlighting-facebook-s-hypocritical-community-guidelines-a6958696.html.

Sinders, Caroline. "Technology Alone Can't Stop Online Harassment." *Vice* (blog), May 28, 2017. https://www.vice.com/en/article/595ed3/technology-alone-cant-stop-online-harassment.

Sung, Morgan. "Andrew Tate Banned from YouTube, TikTok, Facebook and Instagram." *NBC News*, August 22, 2022. https://www.nbcnews.com/pop-culture/viral/andrew-tate-facebook-instagram-ban-meta-rcna43998.

Suzor, Nicolas, Molly Dragiewicz, Bridget Harris, Rosalie Gillett, Jean Burgess and Tess Van Geelen. "Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online." *Policy & Internet* 11, no 1 (2019): 84–103. https://doi.org/10.1002/poi3.185.

Taylor, Josh. "We Unleashed Facebook and Instagram's Algorithms on Blank Accounts. They Served up Sexism and Misogyny." *The Guardian*, July 20, 2024. https://www.theguardian.com/technology/article/2024/jul/21/we-unleashed-facebook-and-instagrams-algorithms-on-blank-accounts-they-served-up-sexism-and-misogyny.

Trott, Verity, Jennifer Beckett and Venessa Paech. "Operationalising 'Toxicity' in the Manosphere: Automation, Platform Governance and Community Health." *Convergence* 28, no 6 (2022): 1754–1769. https://doi.org/10.1177/13548565221111075.

Vaccaro, Kristen, Christian Sandvig and Karrie Karahalios. "'At the End of the Day Facebook Does What It Wants': How Users Experience Contesting Algorithmic Content Moderation." *Proceedings of the ACM on Human-Computer Interaction* 4, no CSCW2 (2020): 167:1–167:22. https://doi.org/10.1145/3415238.

Van Loon, Joost. "Technological Sensibilities and the Cyberpolitics of Gender: Donna Haraway's Postmodern Feminism." *Innovation: The European Journal of Social Science Research* 9, no 2 (1996): 231–43. http://dx.doi.org/10.1080/13511610.1996.9968486.

Wajcman, Judy. *Feminism Confronts Technology*. Philadelphia: Pennsylvania State University Press, 1991.

Wajcman, Judy. *Technofeminism*. Cambridge: Polity Press, 2004.

Wajcman, Judy and Erin Young. "Feminism Confronts AI: The Gender Relations of Digitalisation." In *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, edited by Jude Browne, Stephen Cave, Eleanor Drage and Kerry McInerney, 47–64. Oxford: Oxford University Press, 2023. https://doi.org/10.1093/oso/9780192889898.003.0004.

Weng, Lilian, Vik Goel and Andrea Vallone. "Using GPT-4 for Content Moderation," August 15, 2023. https://openai.com/index/using-gpt-4-for-content-moderation.

Young, Erin, Judy Wajcman and Laila Sprejer. "Mind the Gender Gap: Inequalities in the Emergent Professions of Artificial Intelligence (AI) and Data Science." *New Technology, Work and Employment* 38, no 3 (2023): 391–414. https://doi.org/10.1111/ntwe.12278.