The Politics of Algorithmic Censorship: Automated Moderation and its Regulation

Rachel Griffin

The Politics of Algorithmic Censorship: Automated Moderation and its Regulation

Rachel Griffin, Sciences Po Law School

Forthcoming in *Music and the Politics of Censorship: From the Fascist Era to the Digital Age*, edited by James Garratt, Turnhout, Belgium, Brepols, 2025.

Abstract:

As social media have become vital channels for all kinds of political communication and cultural production, the regulation of social media content has also become a central arena for the contemporary politics of censorship. This chapter offers an overview of the functioning of *algorithmic content moderation*: the automated monitoring and filtering of user-generated content by online platforms, guided by their legal obligations, in-house content policies and commercial objectives. Given the concentrated power of a small number of multinational social media companies, many scholars consider that their *de facto* power to regulate users' speech represents a new and concerning form of private censorship. This power is also routinely leveraged by state authorities, through both formal legal obligations and various forms of more informal cooperation and influence.

In this context, the chapter argues that public and private censorship cannot be clearly distinguished. Ultimately, all moderation of user content plays out against the backdrop of complex entanglements between public institutions and private companies. As such, moderation is always – more or less directly – influenced by public policy, but also by platform companies' commercial objectives. To illustrate these dynamics, the chapter uses a well-publicised case in which Meta (owner of Facebook and Instagram) was revealed to be censoring drill music videos following informal requests from UK police. This case is used to illustrate some of the key political concerns raised by contemporary forms of algorithmic censorship – notably including a lack of transparency and accountability mechanisms governing state intervention, and the replication in algorithmic decision-making software of institutional racism and other historic biases against marginalised groups. The chapter concludes by discussing how algorithmic moderation on social media is likely to develop in the future, highlighting some recent regulatory and technological developments which suggest that state and corporate control of online media will continue to intensify.

Author bio:

Rachel Griffin is a PhD candidate and lecturer in law at Sciences Po Law School. Her doctoral research focuses on how EU social media regulation addresses structural social inequalities in online media, informed by perspectives from law and political economy, critical race theory, and queer and feminist legal theory. She has also worked on research projects relating to platform labour, AI regulation, and intersections between platform regulation and climate policy, and has taught courses at Sciences Po on social media regulation and platform governance.

Introduction

London's Metropolitan Police has an entire unit dedicated to a music genre. Probably to no one's great surprise, it is one primarily popular among young, working-class people of colour[1]. « Project Alpha », which in 2022 employed around 30 officers[2], monitors UK drill music, a rap subgenre popular in deprived neighbourhoods of London, whose lyrics often reference local gang rivalries and real or imagined violent incidents.[3] Due to its alleged potential to encourage violent crime, drill music has been a frequent target of police intervention. This ranges from the use of drill lyrics in criminal trials as evidence of gang affiliations and violent behaviour, to interventions by armed police in the filming of music videos, and even injunctions banning artists from performing certain songs[4]. However, much of Project Alpha's activity also plays out online. Officers monitor the social media accounts of drill artists perceived as dangerous, and request that platforms remove tracks they believe could encourage violence[5].

In 2022, this practice was brought to international attention as the subject of a decision by the Oversight Board[6] – an operationally independent body established by Meta, parent company of Facebook, Instagram and WhatsApp, to advise on how the company regulates users' social media content[7]. At the Met's request, Instagram had removed a music video by drill artist Chinx (OS). The Board concluded that this inadequately respected users' freedom of expression, especially in light of the track's artistic nature, the tenuous evidence that it constituted a threat of violence, and the lack of transparency around Meta's cooperation with police[8]. Perhaps more interestingly[9], the Board's decision also described Meta's decision-making process in detail. It revealed that Meta had originally removed the video in response to an informal email from the Met, and that 164 similar posts were subsequently removed, mostly through an automated system used to automatically scan for copies of previously-removed content. This was one of 992 Met Police requests for social media platforms to remove music content between June 2021 and May 2022,[10] 100% of which related to drill music[11]. Platforms had complied in 89% of cases[12].

As this case illustrates, social media have become a key arena for the contemporary politics of censorship. Cultural production, media consumption and interpersonal communication are increasingly intermediated by a relatively small number of dominant online platforms[13], which now represent key points of control over media and communications[14]. Content moderation – processes by which platforms identify and remove, or otherwise address, content deemed illegal, harmful or undesirable – is widely recognised as essential to create

---

[1] For background on the long history of institutional racism in the Met see TRILLING 2023
[2] ECONOMIST 2022
[3] JALLOH 2023
[4] FATSIS 2019; KEENAN 2021; QUINN et al. 2022 ; JALLOH 2023
[5] ECONOMIST 2022
[6] OVERSIGHT BOARD 2022A
[7] KLONICK 2020
[8] OVERSIGHT BOARD 2022A
[9] Many commentators have questioned the Board's capacities to effectively hold Meta accountable, in particular because its remit and goals are set by Meta. However, one clearly positive outcome is its access to and publication of otherwise non-public information about how Meta's moderation systems work. See DOUEK 2024
[10] OVERSIGHT BOARD 2022B
[11] *Ibid.*
[12] *Ibid.*
[13] SRNICEK 2016; POELL ET AL., 2022
[14] GORWA 2019

usable and safe online spaces[15]. Yet many scholars and activists are concerned about the power of the multinational conglomerates that control today's dominant platforms – including social media, but also other powerful intermediaries like search engines and app stores – to regulate media and communications in accordance with their own commercial interests[16]. As Project Alpha illustrates, these private systems of speech regulation can also be leveraged by state authorities for their own purposes[17].

These concerns are complicated further by platforms' increasing capacities to moderate content automatically, rather than relying on human workers to manually review content[18]. Automated monitoring enables far more effective and pervasive corporate control over what can be said online: all user content and communications can be scanned before they are uploaded, and undesirable behaviour pre-empted before it even takes place[19]. It creates new possibilities for discriminatory censorship[20], given that algorithmic decision-making software has well-studied tendencies to replicate, intensify and scale up existing patterns of social inequality[21]. As this chapter will show, these new, efficient tools of surveillance and control also offer new possibilities for state interference.

This chapter offers a brief introduction to algorithmic content moderation as a 21st-century channel for censorship. This term has most traditionally been understood as referring to regulation of speech by state authorities[22]. However, in an age where companies like Meta, Google and Apple effectively regulate billions of users, and control essential chokepoints to reach public attention, many scholars and activists consider that they exercise power akin to governments and that their moderation practices raise similar normative concerns[23]. Importantly, however, as the chapter will show, presenting corporate content moderation as comparable to but distinct from state censorship does not accurately reflect the institutional reality[24]. Ultimately, commercial content moderation always takes place « in the shadow of the law »[25], influenced though not wholly determined by state policies[26]. There is a spectrum of state involvement encompassing direct legal mandates to censor specific content, informal requests like Project Alpha's, and various more diffuse and indirect forms of influence.

After introducing the basic functioning of algorithmic moderation, the chapter explores some key normative concerns it raises[27]. The Oversight Board's decision on the Chinx (OS) drill video is used as a case study which usefully illustrates not only some typical technical and institutional processes involved in content moderation, but also the complex entanglements of state and corporate power that are at play. The chapter concludes by discussing the potential

[15] ROBERTS 2018A; DOUEK 2021
[16] COBBE 2020; YORK 2021
[17] HINTZ 2016; FROSIO & HUSOVEC 2021; BLOCH-WEHBA 2021
[18] GORWA et al. 2020; COBBE 2020
[19] COBBE 2020
[20] GORWA et al. 2020
[21] For a critical overview of relevant literature see BALAYN & GÜRSES 2021
[22] See OXFORD REFERENCE N.D. (referring to « any regime or context » of control over speech, but emphasising « official grounds » and « governments » as typical contexts).
[23] For a critical overview see GRIFFIN 2023A.
[24] *Ibid.*
[25] KENNEDY 1991, p. 354
[26] GRIFFIN 2023A
[27] Due to space constraints and the author's professional expertise, this discussion focuses primarily on the UK and European contexts.

future outlook for algorithmic censorship, which in an era of « platform capitalism »[28] will surely remain a key area of political intervention and contestation.

Algorithmic content moderation

Content moderation has historically been understood as referring to the policies and processes by which platforms decide whether to remove user content[29]. As platforms' moderation operations have expanded in size and complexity, alongside the development of a professional « trust and safety » community[30] and the burgeoning academic field of « platform governance »[31], moderation is now increasingly understood as encompassing a range of other possible interventions to address harmful or undesirable content. This can include for example labelling possible misinformation with fact-checking information, recommending content to fewer users, or making content and accounts harder to find without removing them entirely[32]. In principle these interventions are based on public « community guidelines » setting out what users are or are not allowed to post – though in practice such guidelines offer a general, incomplete, and often actively misleading picture of the internal rulebooks, informal practices and selective standards that actually determine how content is moderated[33].

Historically, moderation was primarily undertaken by large teams of workers, typically employed through outsourcing companies, often in Global South countries with lower labour costs. Their poor working conditions and the highly stressful nature of their work have been the subject of extensive academic research and critique[34], as well as several lawsuits[35]. Their work remains central to platforms' operations. However, due to several factors – including technological advances, regulatory pressures to moderate content faster and more comprehensively, and commercial pressures to do so in a cost-effective and scalable way – moderation is also increasingly now undertaken through automated filtering software[36].

Two main forms of algorithmic moderation can be distinguished: hash-matching and AI classifiers. Hash-matching is a cryptographic tool used to automatically remove exact or near-exact copies of content previously identified by moderators. It works by creating a unique « hash code » for a given file, typically an image or video; later, (near-)identical copies will produce the same code, so they can efficiently be matched to databases of previously-removed content[37]. Hash-matching tools have prominently been used to address child sexual abuse material and terrorism-related content. In both cases, platforms have commercial and regulatory incentives to cooperate to effectively police such content: they are high priorities for policymakers and law enforcement, they bring significant reputational and commercial risks, and it is widely accepted that they should never be allowed on any platform. This has

---

[28] SRNICEK 2016
[29] ROBERTS 2018A; GILLESPIE 2018
[30] See e.g. the Trust & Safety Professional Association, founded in the US in 2020. There is also an academic *Journal of Online Trust and Safety*, which published its first issue in 2021, and a yearly Trust and Safety Research Conference, hosted at Stanford University.
[31] GORWA 2019
[32] GOLDMAN 2021
[33] GILLESPIE 2018. For example, in 2017 the *Guardian* published leaked excerpts from Facebook's internal rulebook for moderators, which differed significantly from its public-facing community guidelines: HOPKINS 2017. In 2021, further leaks revealed that Meta operated a secret policy that effectively completely exempted certain high-profile users from its official moderation rules: HORWITZ 2021
[34] ROBERTS 2018A; JEREZA 2021; AHMAD 2022
[35] See e.g. PAUL 2020; KIMEU 2023
[36] GORWA et al. 2019
[37] *Ibid.*

led to the development of industry-wide databases, which mean that content removed by one platform can be automatically identified and removed across all participating platforms[38].

Platforms have also relied on hash matching for other content types, including content that is not necessarily illegal, but merely banned under their in-house policies – as happened in the Chinx (OS) case. The Oversight Board's decision revealed that after the video was initially removed in response to the Met's email, it was added to Meta's « Media Matching Service », an in-house tool which stores hash codes for content removed under its rules on « dangerous individuals and organisations », abuse and nudity[39]. In effect, once a human moderator decides to remove content in a particular instance, hash-matching allows this decision to be « scaled up » across the platform, enforcing the ban on that image or video in a continuous and comprehensive way that would not be possible when relying only on human review.

« Fingerprinting » technologies function similarly to hash-matching, in that they scan content and match it to databases of known files, but are designed to be more sensitive to variations and modifications[40]. These are very widely used to enforce music copyrights. Large platforms have long implemented such systems under commercial agreements with major music labels, which are important business partners.[41] They are now also legally obliged to do so under the EU's 2019 Copyright Directive[42]. Often, instead of removing content that makes use of registered copyright works, fingerprinting systems are used to redirect advertising revenue from the user posting the content to the copyright owner. This practice has attracted criticism, since it may involve transformative uses of copyright material (e.g. remixes or reviews) which do not infringe copyright and represent a user's own copyright-protected creative work[43]. Hash-matching and fingerprinting systems which simply look for copies of existing files are incapable of assessing and respecting these exceptions from copyright protection.

In contrast to hash-matching and fingerprinting, AI classifiers are used to analyse previously-unknown content. AI tools are built on training datasets from which they can learn patterns to apply to new content: for example, learning that certain words and linguistic features are associated with hate speech, or certain visual patterns with nudity[44]. Smaller platforms often outsource software development, purchasing standardised moderation software like Google's Perspective[45], while larger platforms with more technical resources are more likely to build or customise their own tools. To do so, they may rely on existing industry-standard datasets, like the ImageNet corpus of labelled pictures[46]. Alternatively, they can build in-house tools or fine-tune existing models using datasets of their own moderators' previous decisions: for example, a platform could fine-tune a general-purpose language model like OpenAI's GPT series (the basis for the popular chatbot programme ChatGPT) to detect hate speech, by feeding it examples of text that was previously labelled as hate speech by moderators[47].

As this suggests, algorithmic moderation does not operate fully autonomously from human intervention. Hash-matching databases and classifiers are built to reproduce previous

[38] *Ibid.*
[39] OVERSIGHT BOARD 2022C
[40] AUDIBLE MAGIC N.D.
[41] TANG 2023
[42] BRIDY 2020
[43] SENFTLEBEN et al. 2023
[44] MONEA 2022
[45] TALAT 2021
[46] MONEA 2022
[47] NICHOLAS & BHATIA 2023

decisions made by human moderators[48], or evaluations by other workers in the burgeoning industry of AI training data production[49]. Even the most advanced AI tools have limited abilities to reliably assess the meaning of content in context[50], so a significant workforce of moderators remains necessary to review and correct automated content classifications[51], as well as providing fresh training data to improve and evaluate AI classifiers and to update them as platforms' rules and policies evolve[52]. And of course, all moderation tools are designed and maintained by platforms' software engineers, to enforce policies written and updated by « trust and safety » staff. Algorithmic moderation decisions are thus best understood as co-produced by networks of technologies and human actors, rather than entirely automated[53].

Algorithmic moderation as private censorship

Although they extend and build on manual review and monitoring, rather than replacing it, algorithmic moderation tools afford new possibilities that were not technically or operationally feasible when relying primarily on manual moderation. Law and technology scholar Jennifer Cobbe argues that two key characteristics of « algorithmic censorship » raise particular normative concerns. First, it allows moderation to operate more comprehensively and at a larger scale. Human moderators could only ever review a tiny subset of the content posted on platforms with millions or billions of users – typically focusing on content that other users have reported as objectionable[54] – but every single post can be scanned at the point of upload by AI classifiers and hash-matching software. Second, as a consequence, platforms can take a « more active, interventionist approach » to moderation, blocking content before it is even posted rather than waiting for it to garner a reaction from other users[55]. Moreover, for Cobbe, pervasive algorithmic filtering implies not only more comprehensive and effective corporate control over online communications, but a qualitative shift in the nature of this control. Content deemed objectionable will not merely be reviewed and sanctioned for non-compliance with rules: it will be technically impossible to post[56].

For many researchers, activists and civil society organisations, this new form of corporate speech regulation represents a new, crucial arena for political struggles over censorship and freedom of expression[57]. Other scholars place more emphasis on the continuities between the power of today's corporate social media, and the outright censorship and more subtle corporate influence that have always characterised commercial media[58]. However, the increasingly pervasive and opaque forms of control enabled by algorithmic moderation[59], and the unprecedented concentration of this control with a small number of global corporations – most prominently including Meta, Google (which owns YouTube) and TikTok, all with

[48] GORWA et al. 2020
[49] DZIEZA 2023
[50] NICHOLAS & BHATIA 2023
[51] In the EU, Article 20 of the 2022 Digital Services Act requires platforms to review decisions at the request of users, « under the supervision of appropriately qualified staff, and not solely on the basis of automated means ». For a detailed analysis of the level of human review required, see GRIFFIN & STALLMAN 2023
[52] DZIEZA 2023
[53] BELLANOVA & DE GOEDE 2021
[54] GILLESPIE 2018
[55] COBBE 2020, p. 739
[56] *Ibid.*
[57] YORK 2021; JØRGENSEN 2019
[58] PICKARD 2022
[59] COBBE 2020

billions-strong user bases across multiple services – are widely considered to pose new and concerning threats to freedom of expression[60].

Today's dominant social media platforms depend on advertiser funding, and this creates incentives to filter and regulate user speech in ways that are familiar from older advertiser-funded media systems[61] – with concerning implications for the freedom and diversity of the public sphere. For example, platforms' policies on hate speech, abuse and harassment are often operationalised using algorithmic tools applying industry-standard metrics of « toxicity ». Yet this metric is technically defined not according to legal or other normative definitions of abusive behaviour, but as speech likely to make other people leave a conversation[62]. Thus, in a subtle redefinition, policies which nominally aim to protect vulnerable users are operationalised using tools that centre platforms' interests: keeping people engaged with conversations and producing more content and ad revenue.

Moderation policies and practices are also significantly influenced by advertisers' concerns about « brand safety », meaning that running ads alongside controversial or offensive content could create negative associations with their brands[63]. For example, almost all major platforms completely ban not only pornography, but any kind of nudity and content considered sexually suggestive, in large part because most advertisers do not consider it « brand safe ». This has often been applied indiscriminately to content that references LGBTQIA+ identities – whether deliberately, because they are considered potentially offensive to the broad « mainstream » audiences that most advertisers wish to target, or inadvertently, because algorithmic moderation tools often associate keywords like « lesbian » and « bisexual » with their use in commercial porn aimed at heterosexual audiences[64].

Considering Meta's commercial incentives as an advertising company sheds new light on its treatment of UK drill music. Music videos made by popular celebrities with « mainstream » appeal are lucrative vehicles for advertising and drivers of user engagement. For these reasons, researchers have observed that they often seem to escape platforms' otherwise draconian regulation of sexually suggestive content.[65] For example, as highlighted by sociologist and pole dancing influencer Carolina Are, major platforms have not taken action against the music video for Grammy-award-winning rapper Cardi B's single « Money », which features nudity and pole dancing, even though much less suggestive pole dancing content by less well-known users is routinely removed.[66] In contrast, UK drill videos represent a relatively niche rap genre popular among young working-class people of colour, who are not a particularly lucrative audience for advertisers[67]. References to crime, violence, and profanity are likely to be deemed offensive and « brand unsafe »[68] within moderation systems that have been described as implementing a white « respectability politics »[69]. Videos like Chinx (OS)'s thus offer little value to today's dominant advertising-funded platforms.

---

[60] JØRGENSEN 2019
[61] BAKER 1992
[62] TALAT 2021
[63] GRIFFIN 2023B
[64] MONEA 2022
[65] ARE & PAASONEN 2021
[66] ARE 2020
[67] LYNES et al. 2020
[68] These are all included in the advertising industry association GARM's taxonomy of « unsafe » content categories: see GRIFFIN 2023B
[69] TALAT 2021

More broadly, « algorithmic bias » – the tendency of algorithmic decision-making tools to replicate and exacerbate unequal treatment of already-marginalised social groups – is a pervasive problem in content moderation, as in many other fields where algorithmic decision-making tools are used[70]. Moderation systems are trained on data from historical decisions – meaning, typically, snap judgments made under intense time pressure by low-paid and poorly-treated workers, which are likely to be influenced by conscious and unconscious bias[71]. They are built to reflect the worldviews of predominantly white, male, Western executives and developers[72], and the commercial priorities of companies whose advertising-based business models incentivise them to focus on suppressing content most likely to be deemed « objectionable » by advertisers and mainstream Western audiences, rather than the content most harmful to their global user bases[73]. Inevitably, then, these systems reproduce existing stereotypes, biases and social inequalities.

Human moderators' decisions are also undoubtedly inflected by such prejudices and stereotypes, as are the rules and policies they apply. However, algorithmic moderation does not only replicate biases in its training data, but can also amplify them[74], For example, commenting on the Chinx (OS) decision, leading platform regulation scholar Daphne Keller noted that moderators working under intense time pressure are particularly likely to be influenced by (un)conscious biases towards seeing Black men as violent or threatening[75]. In this case, the use of hash-matching means any such bias did not just play out in decisions about individual posts – it led to an immediate blanket ban on sharing Chinx (OS)'s music video, anywhere on Instagram. In turn, AI classifiers trained on past moderation decisions like these will « learn » that videos containing groups of Black men and other lyrical, musical and visual tropes associated with drill music are associated with violence and threats. Such biases will thus be intensified and durably encoded into the architecture of the platform.

Algorithmic moderation and state censorship

Algorithmic moderation thus affords new possibilities for corporations to control online speech in line with their commercial objectives. However, it also represents a channel for the exercise of state power. Influencing how powerful platforms regulate their users – what legal scholar Jack Balkin has termed « new school speech regulation » – offers new possibilities for governments to pursue longstanding policy objectives[76]. These range from surveillance and control of speech deemed extremist or politically subversive[77], to responding to « moral panics » over media content regarded as dangerous to children[78]. State censorship of online media may be particularly obvious in authoritarian countries like China, where platforms are subject to explicit and exacting requirements to censor speech deemed politically objectionable by state authorities[79]. However – as the Met Police's Project Alpha illustrates –

[70] BALAYN & GÜRSES 2021
[71] MONEA 2022; KELLER & ACLU 2022
[72] CHEMALY 2019
[73] DE KEULENAAR et al. 2023; ROBERTS 2018B
[74] MONEA 2022
[75] KELLER & ACLU 2022. See also JALLOH 2023 on how prevalent negative stereotypes about Black men influence popular interpretations of drill music and videos
[76] BALKIN 2014
[77] BLOCH-WEHBA 2022
[78] ORBEN 2020
[79] STOCKMANN & LUO forthcoming

policymakers and law enforcement agencies in more democratic countries have also routinely leveraged the power of dominant social media platforms to regulate online speech[80].

This has often involved hard regulation. The EU's 2022 Digital Services Act serves as a leading example[81] of a legislative framework which aims simultaneously to ensure more comprehensive and accurate moderation of illegal or harmful content, and to protect users' freedom of expression[82]. With the latter goal in mind, platforms are required to transparently publish their moderation policies[83], and any time they moderate an item of content – which includes interventions short of removing it entirely, like making it less visible to other users – they must inform the user involved, explaining the reason for the decision and whether it involved algorithmic moderation tools or reporting by other users[84]. Users are entitled to demand a review of the decision, and to appeal to out-of-court dispute resolution institutions[85]. These new protections for users, accompanied by provisions stipulating that content moderation should be objective and proportionate way and should have regard to users' fundamental rights, are intended to avoid arbitrary or discriminatory censorship.

However, scholars have raised doubts about the capacities of these vague, aspirational principles and individualistic procedural rights to address the pervasive and opaque forms of corporate control enabled by algorithmic moderation[86]. And at the same time, other DSA provisions and EU measures (such as the 2019 Copyright Directive and 2021 Terrorist Content Regulation) require or strongly incentivise platforms to expand automated moderation as a way of addressing issues that governments regard as policy priorities, such as extremist content[87]. Overall, the DSA could be understood as a kind of double movement, in which the EU encourages the continued expansion, standardisation and refinement of algorithmic moderation systems, while employing fundamental rights and other safeguards to prevent some of the most egregious forms of arbitrary or discriminatory censorship that they might produce – ultimately helping to institutionalise and legitimise them.

State intervention in platform governance has also frequently involved more indirect or informal channels, which continue to coexist with formal regulation. This can take several forms. As in the example of the Met's Project Alpha, public authorities may wish to effect the removal of specific items of content they deem illegal or (potentially) harmful. Courts or other competent public authorities can order platforms to remove illegal content (a process which is also formalised and subjected to procedural safeguards in the DSA, though the relevant substantive and procedural rules will vary by country[88]). Alternatively, public authorities may informally request, rather than ordering, content removal – either reporting content using the same interfaces available to all platform users, or using informal channels such as emailing personal contacts, as in the Chinx (OS) case. Designated police teams now exist in many countries (including the UK, Israel and several EU member states, as well as Europol) that focus on monitoring and reporting specific types of social media content[89]: often terrorism-

[80] YANG 2023
[81] Numerous other jurisdictions around the world, including the UK, have passed or proposed new legislation with similar aims: for examples see CENTRE FOR DIGITAL WELLBEING 2021
[82] GRIFFIN 2022
[83] Digital Services Act, Article 14
[84] Digital Services Act, Article 17
[85] Digital Services Act, Articles 20-21
[86] GRIFFIN 2022, 2023A
[87] BARATA 2021
[88] Digital Services Act, Article 9
[89] CHANG 2018; BLOCH-WEHBA 2022

related, but also including areas such as undocumented migration[90] and Covid misinformation[91], and of course in the case of the UK rap music.

Importantly, these requests do not generally need to clearly demonstrate the content's illegality. They may only claim that the content violates the platforms' in-house content policies, and suggest that the platform may therefore wish to remove it.[92] As illustrated by the Oversight Board's freedom of information request – which revealed that 89% of Project Alpha's drill-related requests resulted in removal[93] – these suggestions are generally likely to be followed, even if they are not particularly well substantiated. Platforms not only face legal risks if they are later determined to have continued hosting illegal content after it was brought to their attention,[94] but generally benefit from maintaining good relationships with law enforcement and politicians and being seen to take swift action on harmful content.

As well as interventions targeting specific items of content, state authorities can use various forms of soft power and informal influence to shape platforms' general moderation policies and practices. For example, policymakers may encourage and shape the content of nominally voluntary self-regulatory efforts by platforms. Examples include the « Christchurch Call » on terrorist content spearheaded by then-New Zealand Prime Minister Jacinda Ardern and French President Emmanuel Macron[95], or the EU's codes on hate speech and disinformation, which involved active participation and guidance from the European Commission[96]. Policymakers may also communicate informally with platforms or publicly demand that they do more to address harmful content, aiming to influence their policies using reputational pressure, personal relationships, and (more or less implicit) threats of future, harder regulation[97]. For example, during urban riots protesting racist police violence in summer 2023, French government ministers both met privately with executives from major platforms, and made several public statements to the effect that platforms should do more to suppress content encouraging violence or property damage[98].

Finally, even in the absence of active state intervention on particular issues, all commercial content moderation arguably involves some level of state influence operating in the background. Due to several factors – including reputational pressures, vague legal obligations which accord significant discretion to regulators (like the DSA's requirements to take vaguely specified actions against « systemic risks »[99]), and the ever-present threat of stricter regulation if policymakers are unsatisfied – « anticipatory obedience » which ensures moderation practices are broadly aligned with government objectives is generally a good commercial strategy, at least in valuable markets like the EU and US[100]. Platforms' policy staff often have close relationships with politicians, and indeed this may be a key reason they are hired[101].

---

[90] CHANG 2018
[91] BIG BROTHER WATCH 2023
[92] *Ibid.*
[93] OVERSIGHT BOARD 2022B
[94] Under Article 6 of the Digital Services Act, hosting services can be liable for disseminating illegal user-generated content if they do not remove it « expeditiously » when they have knowledge of the illegality.
[95] DOUEK 2019
[96] KELLER 2019
[97] *Ibid.*
[98] GRISON 2023
[99] Digital Services Act, Articles 34-35
[100] KELLER 2019, p. 2
[101] SWENEY 2018; WOFFORD 2022

These power dynamics and ongoing relationships with policymakers are always operating in the background to influence moderation practices. Policies will be written with legal risks and government relations in mind, and may draw on government sources: for example, Meta's list of « dangerous individuals and organisations », banned and policed through algorithmic moderation tools including its Media Matching Service, is drawn from the US government's list of banned terrorist organisations[102]. State intervention at a given point in time will also continue to exert an influence on future moderation practices, through algorithmic tools which reproduce or learn from past decisions.

Overall, through all these coexisting and interrelated channels of influence, it is clear that the « private ordering » systems through which platform companies regulate their users are pervasively shaped by the influence of state actors, and offer powerful new channels for state censorship. Algorithmic moderation, in particular, enables far more granular, pervasive and pre-emptive interventions than traditional legal sanctions. Censorship regimes that subject publications to controls before they can be published – as is now the case for all social media content subject to state-mandated algorithmic filtering – have traditionally been regarded as particularly threatening to freedom of expression, associated with authoritarian states and historic systems of media regulation that have now been largely abandoned by Western democracies[103]. Informally pressuring private actors to regulate content « voluntarily » – described as « laundering » of state censorship by legal scholar Hannah Bloch-Wehba[104] – also evades many of the legal safeguards and other accountability mechanisms, such as public scrutiny, that would apply to official state interventions[105].

It would go too far to say that any influence or cooperation between state actors and platforms is undesirable: it may often be justified to prevent serious illegal activity, such as the dissemination of child sexual abuse material, or to hold platforms accountable to public-interest goals, like preventing abuse and harassment. At the same time, however, the various mechanisms of state influence identified above raise serious concerns about the suppression of political dissent and unpopular minorities. This is not only due to deliberate, targeted abuses of power by state authorities, but also because of inevitable issues of algorithmic bias and discrimination, as discussed in the previous section. Another particularly concerning aspect is the often informal, covert and unaccountable nature of these interventions. For example, the monitoring and reporting practices of the Israeli police's « Cyber Unit » were recently challenged in a lawsuit that made it all the way to the Supreme Court, but which was ultimately dismissed on the basis that the claimants could not definitively prove that any particular moderation decision had involved state intervention.[106]

Project Alpha offers a vivid example of how, in the absence of strong accountability mechanisms, state power to censor online speech is likely to be abused. The Met Police's targeting of drill shows continuities with longer histories of police racism and tendencies to stigmatise « young urban black men – and the forms of culture that appear tied to this population – [as] a threat to the civic mainstream »[107] despite the very tenuous evidence for links between drill music and violent crime[108]. This has real consequences for the freedom of

[102] BIDDLE 2021
[103] BARENDT 2007, ch. 4
[104] BLOCH-WEHBA 2022, p. 1302
[105] KELLER 2019
[106] KELLER 2022
[107] LYNES et al. 2020, p. 1202
[108] FATSIS 2019; ECONOMIST 2022

artistic expression of the users involved, and the communities to which they belong.[109] An in-depth journalistic investigation of the UK drill scene by *Vice* argues that police interventions have done « an exceptional job of eradicating the chances for aspiring rappers to transition toward full-time music: Debut singles and sophomore offerings disappear from YouTube, with potential up-and-comers' careers obliterated before they begin. »[110] State censorship, « laundered » through corporate moderation systems which evaluate drill music as having little commercial value, suppresses the creativity, self-expression and career aspirations of young people to whom UK society otherwise offers few opportunities and resources[111].

Monitoring of social media content by police and private companies also feeds into broader systems of surveillance and social control[112]. In the UK, drill music content has often been used in trials as evidence of gang associations and criminal activity. Police and prosecutors use rap lyrics and videos to evoke negative stereotypes about Black culture, reinforcing racist disparities within the justice system[113]. For sociologist Lambros Fatsis, stigmatising and policing drill music helps uphold the UK's broader architecture of racial and class inequality, as « the neoliberal state accuses residents for the deterioration in their surroundings; often attributing such decline to a lack of civility and a cultural propensity for gang violence. »[114] Similar alignments between platform policies and state ideologies can be observed in other contexts. For example, the use of state terrorism blocklists to identify « dangerous individuals and organisations » means that platforms' moderation systems replicate the Islamophobic policies of state counterterrorism agencies[115]. Experts have also linked platforms' close cooperation with Israeli authorities, including the Cyber Unit, to the repeated and ongoing suppression of social media content relating to pro-Palestinian political activism[116].

Algorithmic censorship: the future outlook

At this point, algorithmic moderation is an embedded and essential feature of online environments. It is necessary to create usable online spaces that are not quickly overwhelmed with spam and genuinely harmful content, such as the abuse, harassment and hate speech that have historically denied members of marginalised social groups equal access to public spaces[117]. It is also integral to the business models of corporate platforms that need to quickly suppress « objectionable » content, in order to keep users engaged and prove their « brand safety » to advertisers[118]. And finally, it offers state authorities powerful new tools to monitor and control types of speech deemed threatening – tools they are unlikely to give up. Looking ahead, it seems likely that the role of algorithmic censorship in modulating online speech will only increase. Here, three important trends can be tentatively identified.

First, AI moderation tools are improving fast. Historically, such tools had very limited ability to analyse the meaning of speech in context; to understand nuances or implicit meanings evident to human audiences; or to analyse mixed media content like videos or text-and-image

---

[109] JALLOH 2023
[110] HODGSON 2023
[111] LYNES et al. 2020; THAPAR 2023
[112] BLOCH-WEHBA 2022
[113] KEENAN 2021; QUINN et al. 2022; JALLOH 2023
[114] FATSIS 2019, p. 2
[115] BIDDLE 2021
[116] HUMAN RIGHTS WATCH 2021
[117] FRANKS 2021
[118] GILLESPIE 2018; DE KEULENAAR et al. 2023; GRIFFIN 2023B

memes[119]. However, at the time of writing in mid-2023, we are seeing rapid advances in the development and commercialisation of « large language models » (LLMs), like those behind ChatGPT. Trained on vast text corpora, they have markedly improved abilities to analyse and respond to text input, and can be effectively « fine-tuned » for specialist tasks (like enforcing specific content policies) using relatively little additional data[120]. Multimodal LLMs which can analyse visual and mixed media content are also seeing rapid improvements[121].

Such systems' capabilities should not be overstated: their performance is far worse outside of English and a few other « high-resource languages »[122], and they are still nowhere close to achieving human-level comprehension[123]. In any case, the inherent indeterminacy and contestability of content policies – what is « hateful » or « sexually suggestive » is inevitably subjective – means that perfectly accurate enforcement is not just technically out of reach, but intrinsically impossible. Nonetheless, the reliability of AI moderation tools and the range of criteria they can assess will likely continue improving rapidly.

This has many advantages, such as reducing the incidence of arbitrary and mistaken content removals – which tend to disproportionately impact marginalised groups[124] – and more effectively protecting users against harmful content. At the same time, it raises the prospect of ever more granular and pervasive corporate control over online speech[125]. Such control is easily abused for political purposes, and its exercise will ultimately be driven by platforms' commercial incentives, which may sometimes overlap with but do not inherently correspond to these public-interest goals.

Second, and relatedly, the scope of application of automated moderation tools will likely continue to expand. We have already seen a « function creep » where tools are used for more and more types of content and policy areas: for example, the Oversight Board's decisions show that Meta (and likely other leading platforms) now use hash-matching tools originally developed to address serious forms of illegal content, like child sexual abuse material, to efficiently scale up the moderation of content which may not even be illegal[126]. This expansion seems likely to continue, for multiple reasons.

Technical advances are one relevant factor. Given their efficiency and scalability, increasingly capable and accurate AI classifiers will likely be deployed in more situations previously considered to require human assessment, even if they remain short of human-level performance. Moreover, algorithmic moderation tools are not just developed to implement pre-existing policies: the available tools also shape what kinds of policies are possible or practical[127]. The capacity to filter content instantaneously and pre-emptively across an entire platform – not only by searching for exact matches of previously-removed content or using relatively crude criteria like blanket bans on nudity, but with AI classifiers which can make increasingly nuanced and context-sensitive assessments of meaning – may open the door to entirely new content policies and enforcement practices.

---

[119] DUARTE & LLANSÓ 2017
[120] NICHOLAS & BHATIA 2023
[121] YIN et al. 2023
[122] NICHOLAS & BHATIA 2023
[123] BENDER & KOLLER 2020
[124] GRIFFIN 2022
[125] COBBE 2020
[126] OVERSIGHT BOARD 2022A, 2022C
[127] GILLESPIE 2018

The other key factor is regulatory pressure. Governments have varying capacities to effectively regulate the enormously wealthy US-based multinational companies that control leading platforms – often depending largely on how valuable their consumers are to platforms as consumers and advertising targets[128]. However, there is now concerted pressure coming from a number of influential jurisdictions and multinational coalitions to expand the scope of automated moderation. Notably, there are regulatory proposals in the EU and UK to expand automated moderation to entirely new areas like private messaging, through tools like « client-side scanning », which allows encrypted messages to be monitored using software running on personal devices[129].

In the social media context, platforms have already effectively been required to scale up automated moderation by existing measures like the EU's DSA and Terrorist Content Regulation, the UK's proposed Online Safety Bill, and similar legislation in other jurisdictions which impose broadly-defined duties for platforms to take proactive action against various types of harms and risks. Companies generally enjoy significant discretion over how to fulfil these legal duties, yet it seems inevitable that expanding automated moderation will be their main response. Not only is it cheaper and more scalable than, for example, hiring more trust and safety staff; it is also legally attractive. Because such tools are already in such widespread use, platform companies can argue that they represent industry-standard best practices. For big tech companies, the idea that technological advances like LLMs present efficient, effective technological solutions is ideologically and commercially appealing[130]; for smaller platforms with limited resources, purchasing off-the-shelf moderation software from larger companies represents a practical and relatively affordable way to deal with commercial and legal demands for effective moderation.

Third and finally, shifting trends in the functionalities and use of social media also suggest that new forms of algorithmic intervention will continue to emerge. For example, popular platforms have increasingly shifted away from displaying content from people users already know or follow, and towards using algorithmic recommendation systems to surface new and unknown content[131]. In this context, « visibility moderation » – recommending content less prominently or otherwise making it less accessible to other users, rather than removing it entirely – has become increasingly central to platforms' management of legal, reputational and commercial risks[132]. While this can in some ways represent a more proportionate response to harmful content which is less restrictive of free speech and access to information than removing content entirely – and can certainly be presented in that way by platforms – it also raises difficult questions about the legitimate uses and potential abuses of platforms' power to shape online discourse by determining what speech can reach an audience[133].

Major social media platforms, search engines and other online tools are also increasingly developing features based on generative AI, like chatbots and image-generation tools[134]. This will present similar questions about the exercise of power through content moderation. The work of human moderators, and AI models built to replicate their evaluations of whether AI-generated content is safe, appropriate or helpful, has been central in making generative AI

[128] DE GREGORIO & STREMLAU, 2023
[129] GORWA 2022
[130] MOROZOV 2013
[131] NARAYANAN 2023
[132] ZENG & KAYE 2022; GILLESPIE 2022
[133] GILLESPIE 2022
[134] CHOW & PERRIGO 2023; REUTERS 2023

tools commercially viable[135]. Ultimately, the algorithmic tools that determine what kind of content should be produced and what should be blocked as unsafe or inappropriate will ultimately reflect the commercial priorities of the companies that develop them – and they will likely become a new target for state intervention, both legal and informal. As AI-generated media become ubiquitous online, ever more of what we see in the media may be shaped by pervasive state and corporate control.

French philosopher Michel Foucault, known for his seminal and sombre work on surveillance, discipline and power, stated in an interview near the end of his life that, « My point is not that everything is bad, but that everything is dangerous. »[136] The position that any and all state censorship is bad is rightly regarded as extremist – especially by feminist and antiracist scholars who emphasise the dangers that unrestricted speech (on- and offline) has historically posed for minority groups[137]. Content moderation, including algorithmic moderation, is essential to create safe and inclusive spaces for online media and communication. Yet the new channels for state and corporate power outlined in this chapter pose very real dangers: suppression of political dissent, systemic discrimination, and the channelling of all online media and communication in line with the priorities of corporate advertisers, to name but a few. Current trends suggest that the role of algorithmic censorship in online media and civic life will only continue to expand. Difficult questions lie ahead about when and how it can be used legitimately, and what safeguards might prevent the worst abuses of power.

[135] DZIEZA 2023; WONG 2023
[136] BERNSTEIN 1994, p. 226
[137] FRANKS 2022

Bibliography

AHMAD, Sana. 'Who moderates my social media? Locating Indian workers in the global content moderation practices', in: *Challenges and perspectives of hate speech research*, edited by Christian Strippel et al., Berlin, Boehland & Schremmer Verlag, 2023 (Digital Communication Research), pp. 111-25.

ARE, Carolina. 'New Terms of Use On Instagram: The End of Nudity?', *Blogger On Pole* (30 November 2020) <https://bloggeronpole.com/2020/11/terms-of-use/>.

ARE, Carolina and PAASONEN, Susanna. 'Sex in the shadows of celebrity', in: *Porn Studies*, 8/1 (2021), pp. 411-19.

AUDIBLE MAGIC. 'Technology', Audible Magic (n.d.) <https://www.audiblemagic.com/technology/>.

BALAYN, Agathe and GÜRSES, Seda. 'Beyond Debiasing: Regulating AI and its inequalities', EDRi (21 September 2021), <https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/>.

BALKIN, Jack. 'Old-School/New-School Speech Regulation', in: *Harvard Law Review*, 127/8 (2014), pp. 2296-342.

BAKER, C. Edwin. 'Advertising and a Democratic Press', in: *University of Pennsylvania Law Review*, 140/6 (1992), pp. 2097-243.

BARATA, Joan. 'The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations', *DSA Observatory* (27 July 2021) <https://dsa-observatory.eu/2021/07/27/the-digital-services-act-and-its-impact-on-the-right-to-freedom-of-expression-special-focus-on-risk-mitigation-obligations/>.

BARENDT, Eric. *Freedom of Speech* (2nd edn), Oxford, Oxford University Press, 2007.

BELLANOVA, Rocco and DE GOEDE, Marieke. 'Co-Producing Security: Platform Content Moderation and European Security Integration', in: *Journal of Common Market Studies*, 60/5 (2021), pp. 1316-34.

BENDER, Emily and KOLLER, Alexander. 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data', in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 5185-98.

BERNSTEIN, Richard. 'Foucault: Critique as a Philosophical Ethos', in: *Critique and Power: Recasting the Foucault/Habermas Debate*, edited by Michael Kelly, Cambridge, Massachusetts, MIT Press, 1994, pp. 211-42.

BIDDLE, Sam. 'Revealed: Facebook's Secret Blacklist of "Dangerous Individuals and Organizations"', *The Intercept* (12 October 2021) <https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/>.

BIG BROTHER WATCH. 'Ministry of Truth: The Secretive Government Units Spying On Your Speech', Big Brother Watch (January 2023) <https://bigbrotherwatch.org.uk/wp-content/uploads/2023/01/Ministry-of-Truth-Big-Brother-Watch-290123.pdf>.

BLOCH-WEHBA, Hannah. 'Content Moderation as Surveillance', in: *Berkeley Technology Law Journal*, 36 (2021), pp. 1297-340.

BRIDY, Annemarie. 'The Price of Closing the Value Gap: How the Music Industry Hacked EU Copyright Reform', in: *Vanderbilt Journal of Entertainment and Technology Law*, 22 (2020), pp. 323-58.

CENTRE FOR DIGITAL WELLBEING. 'International Regulation of Social Media', Centre for Digital Wellbeing (December 2021) <https://digitalwellbeing.org.au/wp-content/uploads/2021/12/International-Regulation-of-Social-Media.pdf>.

CHANG, Brian. 'From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU', in: *Columbia Human Rights Law Review*, 49/2 (2018), pp. 114-212.

CHEMALY, Soraya. 'Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the Silencing of Marginalized Voices' in: *Free Speech in the Digital Age*, edited by Susan J. Brison and Katharine Gelber, Oxford, Oxford University Press, 2019, pp. 150–169,

CHOW, Andrew R. & PERRIGO, Billy. 'The AI Arms Race Is Changing Everything', *TIME* (17 February 2023) <https://time.com/6255952/ai-impact-chatgpt-microsoft-google/>.

COBBE, Jennifer. 'Algorithmic Censorship by Social Platforms: Power and Resistance', in: *Philosophy & Technology*, 34 (2021), pp. 739-66.

DE GREGORIO, Giovanni and STREMLAU, Nicole. 'Inequalities in content moderation', in: *Global Policy* (2023).

DE KEULENAAR, Emillie, MAGALHÃES, João C. and GANESH, Bharath. 'Modulating moderation: a history of objectionability in Twitter moderation practices', in: *Journal of Communication*, 73/3 (2023), pp. 273-87.

DOUEK, Evelyn. 'Two Calls for Tech Regulation: The French Government Report and the Christchurch Call', *Lawfare* (18 May 2019), <https://www.lawfaremedia.org/article/two-calls-tech-regulation-french-government-report-and-christchurch-call>.

DOUEK, Evelyn. 'Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability', in: *Columbia Law Review*, 121/3 (2021), pp. 759-834.

DOUEK, Evelyn. 'The Meta Oversight Board and the Empty Promise of Legitimacy', in: *Harvard Journal of Law & Technology*, 37 (forthcoming 2024).

DUARTE, Natasha and LLANSÓ, Emma. 'Mixed Messages? The Limits of Automated Social Media Content Analysis', Center for Democracy & Technology (28 November 2017)

<https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/>.

DZIEZA, Josh. 'AI Is a Lot of Work', *The Verge* (20 June 2023) <https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots>.

THE ECONOMIST. 'The British police unit helping remove drill-music videos from the web', *The Economist* (7 May 2022), <https://www.economist.com/britain/2022/05/07/the-british-police-unit-helping-remove-drill-music-videos-from-the-web>.

FATSIS, Lambros. 'Policing the beats: The criminalisation of UK drill and grime music by the London Metropolitan Police', in: *The Sociological Review*, 67/6 (2019).

FRANKS, Mary Anne. 'Beyond the Public Square: Imagining Digital Democracy', in: *Yale Law Journal Forum*, 131 (2021), pp. 427-53.

FROSIO, Giancarlo and HUSOVEC, Martin. 'Accountability and Responsibility of Online Intermediaries', in: *Oxford Handbook of Online Intermediary Liability*, edited by Giancarlo Frosio, Oxford, Oxford University Press, 2021, pp. 612-30.

GILLESPIE, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, New Haven, Yale University Press, 2018.

GILLESPIE, Tarleton. 'Do Not Recommend? Reduction as a Form of Content Moderation', in: *Social Media + Society* (2022).

GOLDMAN, Eric. 'Content Moderation Remedies', in: *Michigan Technology Law Review*, 28/1, pp. 1-59.

GORWA, Robert. 'What is platform governance?', in: *Information, Communication & Society*, 22/6 (2019), pp. 854-871.

GORWA, Robert. 'European Security Officials Double Down on Automated Moderation and Client-Side Scanning', *Lawfare* (15 June 2022) <https://www.lawfaremedia.org/article/european-security-officials-double-down-automated-moderation-and-client-side-scanning>.

GORWA, Robert, BINNS, Reuben and KATZENBACH, Christian. 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', in: *Big Data & Society* (2020).

GRIFFIN, Rachel. 'The Sanitised Platform', in: *JIPITEC*, 13/1 (2022), pp. 36-52.

GRIFFIN, Rachel. 'Public and private power in social media governance: multistakeholderism, the rule of law and democratic accountability', in: *Transnational Legal Theory*, 14/1 (2023a), pp. 46-89.

GRIFFIN, Rachel. 'From brand safety to suitability: advertisers in platform governance', in: *Internet Policy Review*, 12/3 (2023b).

GRIFFIN, Rachel and STALLMAN, Erik. 'A Systemic Approach to Implementing the DSA's Human-in-the-Loop Requirement', *Verfassungsblog* (forthcoming 2023).

GRISON, Thibault. 'Y a trop de choses qui m'interpellent sur cette ITW de Thierry Breton hier annonçant une modération exceptionnelle des """contenus haineux""" à partir du 25 août sur @franceinfo . [Je voulais poster hier, mais je me suis endormi plus tôt, oups] 🧵', Twitter (11 July 2023) <https://twitter.com/thibo_grison/status/1678674092222824448>.

HINTZ, Arne. 'Restricting digital sites of dissent: commercial social media and free expression', in: *Critical Discourse Studies*, 13/3 (2016), pp. 325-40.

HODGSON, Jaimie. 'The Secret History of Drill', *VICE* (25 July 2023) <https://www.vice.com/en/article/4a3x3m/secret-history-drill-music>.

HOPKINS, Nick. 'Revealed: Facebook's internal rulebook on sex, terrorism and violence', *The Guardian* (21 May 2017) <https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence>.

HORWITZ, Jeff. 'Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.', *Wall Street Journal* (13 September 2021) <https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353>.

HUMAN RIGHTS WATCH. 'Israel/Palestine: Facebook Censors Discussion of Rights Issues', Human Rights Watch (8 October 2021) <https://www.hrw.org/news/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues>.

JALLOH, Tareeq Omar. 'Does the Critical Scrutiny of Drill Constitute an Epistemic Injustice?' in: *The British Journal of Aesthetics*, 62/4, pp. 633-651.

JEREZA, Rae. 'Corporeal moderation: digital labour as affective good', in: *Social Anthropology*, 29/4, pp. 928-943.

JØRGENSEN, Rikke Frank (ed.). *Human Rights in the Age of Platforms*, Cambridge, Massachusetts, MIT Press, 2019.

KEENAN, Maeve. 'JUSTICE report: Report finds misunderstanding of Drill music is leading to unfair convictions', Youth Justice Legal Centre (2021) <https://yjlc.uk/resources/legal-updates/justice-report-report-finds-misunderstanding-drill-music-leading-unfair>.

KELLER, Daphne. 'Who Do You Sue? State and Platform Hybrid Power Over Online Speech', Hoover Institution, Aegis Series Paper No. 1902 (29 January 2019) <https://www.hoover.org/research/who-do-you-sue>.

KELLER, Daphne. 'When Platforms Do the State's Bidding, Who Is Accountable? Not the Government, Says Israel's Supreme Court', *Lawfare* (7 February 2022) <https://www.lawfaremedia.org/article/when-platforms-do-states-bidding-who-accountable-not-government-says-israels-supreme-court>.

KELLER, Daphne and ACLU. 'Daphne Keller and ACLU File Comment to Meta Oversight Board in "UK Drill Music" Case', Stanford Cyber Policy Center (23 August 2022) <https://cyber.fsi.stanford.edu/news/daphne-keller-and-aclu-file-comment-uk-drill-music-case>.

KENNEDY, Duncan. 'The Stakes of Law, or Hale and Foucault!', in: *Legal Studies Forum*, XV/4 (1991), 327-66.

KIMEU, Caroline. 'High-profile lawsuit against Meta can be heard in Kenya, Nairobi court rules', *The Guardian* (7 February 2023) <https://www.theguardian.com/global-development/2023/feb/07/lawsuit-meta-kenya-nairobi-court-rules-facebook-ptsd>.

KLONICK, Kate. 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression', in: *Yale Law Journal*, 129/8 (2020), pp. 2418-99.

LYNES, Adam, KELLY, Craig and KELLY, Emma. 'THUG LIFE: Drill music as a periscope into urban violence in the consumer age', in: *British Journal of Criminology*, 60/5 (2020), pp. 1201-19.

MONEA, Alexander. *The Digital Closet: How the Internet Became Straight.* Cambridge, Massachusetts, MIT Press, 2022.

MOROZOV, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York, Public Affairs, 2013.

NARAYANAN, Arvind. 'Understanding Social Media Recommendation Algorithms', Knight First Amendment Institute (9 March 2023) <https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms>.

NICHOLAS, Gabriel and BHATIA, Aliya. 'Lost in Translation: Large Language Models in Non-English Content Analysis', Center for Democracy & Technology (23 May 2023) <https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/>.

ORBEN, Amy. 'The Sisyphean Cycle of Technology Panics', in: *Perspectives on Psychological Science*, 15/5 (2020).

OVERSIGHT BOARD. 'Oversight Board overturns Meta's decision in "UK drill music" case', Oversight Board (November 2022) <https://www.oversightboard.com/news/413988857616451-oversight-board-overturns-meta-s-decision-in-uk-drill-music-case/>.

OVERSIGHT BOARD. 'January 2023 updated freedom of information request, Metropolitan Police', Oversight Board (November 2022) <https://www.oversightboard.com/news/413988857616451-oversight-board-overturns-meta-s-decision-in-uk-drill-music-case/>.

OVERSIGHT BOARD. 'Colombian police cartoon', Oversight Board (2022) <https://www.oversightboard.com/decision/FB-I964KKM6>.

OXFORD REFERENCE. 'Quick Reference: Censorship', Oxford Reference (n.d.), <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095558166>.

PAUL, Kari. 'Facebook to pay $52m for failing to protect moderators from "horrors" of graphic content', *The Guardian* (12 May 2020) <https://www.theguardian.com/technology/2020/may/12/facebook-settlement-mental-health-moderators>.

PICKARD, Victor. 'The Great Reckoning', Knight First Amendment Institute (24 February 2022) <https://knightcolumbia.org/content/the-great-reckoning>.

POELL, Thomas, NIEBORG, David B. and DUFFY, Brooke Erin. *Platforms and Cultural Production*, Cambridge, Polity, 2021.

QUINN, Eithne, WHITE, Joy and STREET, John. 'Introduction to special issue: *Prosecuting and Policing Rap*', in: *Popular Music*, 41/4 (2022), pp. 419-426.

REUTERS. 'Meta previews generative AI tools planned for its platforms', *Reuters* (8 June 2023) <https://www.reuters.com/technology/meta-previews-generative-ai-chatbot-planned-whatsapp-messenger-company-all-hands-2023-06-08/>.

ROBERTS, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*, New Haven, Yale University Press, 2018a.

ROBERTS, Sarah T. 'Digital detritus: "Error" and the logic of opacity in social media content moderation', in: *First Monday*, 23/3 (2018b).

SENFTLEBEN, Martin, QUINTAIS, João Pedro and MEIRING, Arlette. 'Outsourcing Human Rights Obligations and Concealing Human Rights Deficits: The Example of Monetizing User-Generated Content Under the CDSM Directive and the Digital Services Act', SSRN (28 April 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4421150>.

SRNICEK, Nick. *Platform Capitalism*, Cambridge, Polity, 2016.

STOCKMANN, Daniela and LUO, Ting. *Governing Digital China: Managing Citizen Participation During Political Change*. Cambridge University Press (forthcoming).

SWENEY, Mark. 'Facebook hires Nick Clegg as head of global affairs', *The Guardian* (19 October 2018) <https://www.theguardian.com/technology/2018/oct/19/facebook-hires-nick-clegg-as-head-of-global-affairs>.

TALAT, Zeerak. *'It ain't all good": machinic abuse detection and marginalisation in machine learning*, unpublished Ph.D. Diss., Sheffield, University of Sheffield, 2021 <https://etheses.whiterose.ac.uk/30950/>.

TANG, Xiyin. 'Privatizing Copyright', in: *Michigan Law Review*, 121 (forthcoming 2023).

THAPAR, Ciaran. 'Rap and drill music give voice to the pain of life in a world of violence, and YouTube is the new amphitheatre', *The Guardian* (15 April 2023)

<https://www.theguardian.com/commentisfree/2023/apr/15/rap-drill-music-life-violence-youtube-ancient-athens-catharsis-music-teenagers>.

TRILLING, Daniel. 'Not Much Like Consent', in: *London Review of Books*, 45/7 (2023), <https://www.lrb.co.uk/the-paper/v45/n07/daniel-trilling/not-much-like-consent>.

WOFFORD, Benjamin. 'The Infinite Reach of Joel Kaplan, Facebook's Man in Washington', *Wired* (10 March 2022), <https://www.wired.com/story/facebook-joel-kaplan-washington-political-influence/>.

WONG, Matteo. 'America Already Has an AI Underclass', *The Atlantic* (26 July 2023) <https://www.theatlantic.com/technology/archive/2023/07/ai-chatbot-human-evaluator-feedback/674805/>.

YANG, Fan. 'A Glitch in Translation: (Self-)Orientalism and PostOrientalism in Platform Governance', in: *Yale ISP-WIII Essay Series: Platform Governance Terminologies*, edited by Mehtab Khan and Brianna Yang <https://law.yale.edu/sites/default/files/area/center/isp/documents/translation_ispessayseries_2023.pdf>.

YIN, Shukang, FU, Chaoyou, ZHAO, Sirui, LI, Ke, SUN, Xing, XU, Tong and CHEN, Enhong. 'A Survey on Multimodal Large Language Models', arXiv (23 June 2023) <https://arxiv.org/abs/2306.13549>.

YORK, Jillian C. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*, New York, Verso, 2021.

ZENG, Jing & KAYE, D. Bondy Valdovinos. 'From content moderation to *visibility moderation*: A case study of platform governance on TikTok', in: *Policy & Internet*, 14/1 (2022), pp. 79-95.