White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/157914/

Version: Published Version

---

**Article:**
Gerrard, Y. and Thornham, H. (2020) Content moderation: Social media's sexist assemblages. New Media and Society, 22 (7). pp. 1266-1286. ISSN: 1461-4448

https://doi.org/10.1177/1461444820912540

---

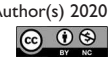
**Reuse**
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial (CC BY-NC) licence. This licence allows you to remix, tweak, and build upon this work non-commercially, and any new works must also acknowledge the authors and be non-commercial. You don't have to license any derivative works on the same terms. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

**Takedown**
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Check for updates

*Article*

new media & society
2020, Vol. 22(7) 1266–1286
© The Author(s) 2020

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1461444820912540
journals.sagepub.com/home/nms


# Content moderation: Social media's sexist assemblages

**Ysabel Gerrard**
University of Sheffield, UK

**Helen Thornham**
University of Leeds, UK

## Abstract

This article proposes 'sexist assemblages' as a way of understanding how the human and mechanical elements that make up social media content moderation *assemble* to perpetuate normative gender roles, particularly white femininities, and to police content related to women and their bodies. It investigates sexist assemblages through three of many potential elements: (1) the normatively gendered content presented to users through in-platform keyword and hashtag searches; (2) social media platforms' community guidelines, which lay out platforms' codes of conduct and reveal biases and subjectivities and (3) the over-simplification of gender identities that is necessary to algorithmically recommend content to users as they move through platforms. By the time the reader finds this article, the elements of the assemblages we identify might have shifted, but we hope the framework remains useful for those aiming to understand the relationship between content moderation and long-standing forms of inequality.

## Keywords

Algorithms, assemblage theory, content moderation, gender, Pinterest, sexism, social media

**Corresponding author:**
Ysabel Gerrard, Department of Sociological Studies, University of Sheffield, Sheffield S10 2TU, South Yorkshire, UK.
Email: y.gerrard@sheffield.ac.uk

## Introduction

Banning images of 'female-presenting nipples' on Tumblr (Duguay, 2018), limiting the results of hashtag searches related to women of colour – like #mixedgirls, #blackgirls and #mexicangirls – on Instagram (Drewe, 2016), and problematising images of underweight female bodies on Pinterest (Gerrard, 2018) are only a few recent examples of the nuanced human and machine policing of social media content related to women and their bodies. In this article, we explore how gender – specifically *sexism*, defined here and elsewhere as the discrimination of a person on the basis of their sex or gender (among others, Douglas, 2010; Gill, 2014) – can be used as an analytic lens through which to understand the underpinning logics, processes and importantly, consequences of social media content moderation. Furthermore, we show how gender, race and other identity markers intersect (Crenshaw, 1989) through the various processes of content moderation to reproduce stereotypes of those who experience eating disorders. Content moderation – 'the organized practice of screening user generated content (UGC) posted to Internet sites, social media and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction' (Roberts, 2017a: 44) – has gained increased scholarly and public attention in recent years. As Gillespie (2010) notes, social media companies have become 'more like traditional media than they care to admit' (p. 359): they are increasingly setting the parameters of 'acceptable' social conduct and, as ever, this is having consequences for society's most marginalised groups. This concern speaks directly to the focus of this *New Media and Society* Special Issue: understanding the social in a digital age.

Researchers have so far focussed on the human labour behind content moderation (Carmi, 2019; Roberts, 2016, 2017b, 2019), social media platforms' changing responsibilities (Gillespie, 2015, 2018; Suzor, 2019), users' experiences of platforms' interventions (Duguay et al., 2018; Gerrard, 2018; Myers-West, 2018) and community-driven forms of moderation (Lo, 2018; Seering et al., 2019; Squirrell, 2019). Uniting this research is a focus on humans and machines, partly through the legacy of Science and Technology Studies (STS) scholarship (e.g. Barad, 2009; Suchman, 2007; Wajcman, 1991) and also because of the increasing need to understand how social norms 'leak across', to use Cheney-Lippold's (2017: 143) term, to content moderation processes and vice versa. A fundamental yet academically under-addressed part of content moderation debates is *gender*; specifically, how gender norms factor into and get reproduced by content moderation processes and outcomes.

This is not to say that there is an absence of scholarship looking at inequalities. Researchers interested in issues of gender have to date explored individual components of what we are calling social media's *sexist assemblages* (and which we describe below). For example, the 'baking' of gender into social media's design (Bivens and Haimson, 2016; Kirchner, 2015), the intersection of gendered and racialised norms through search results – for example, how Google used to recommend a website called hotblackpussy.com as one of the first search results for 'Black girls' (Noble, 2018) – and the gendering of artificial intelligent 'digital assistants' like Amazon's Alexa (Neff, 2018). Our use of these examples in addition to discussing our own also demonstrates how gender, race and other identity markers intersect (Crenshaw, 1989) to discriminate. The sexist assemblages that

we identify in this article are therefore much more than *sexist* assemblages. One of the critiques levelled at such previous work as Bucher (2018) has argued (drawing on the work of scholars like DeLanda, 2006; Law, 2004; Mackenzie, 2006; Suchman, 2007), relates to their exploration of singular or static elements of social media, producing a less sophisticated understanding than is needed of both the object of enquiry (the given social media platform) and the work that algorithms and human interaction do (pp. 54–55). For Bucher (2018), this means that understanding platforms as anything other than a range of permutations, as multiplicity and as *processes* is too limiting (p. 49).

In keeping with Bucher's (2018) argument, we draw on the conception of platforms as *assemblages* (pp. 50–51) to acknowledge the dynamic and iterative processes of platforms always-already coming into being, rather than considering them as static or fixed objects of study. At the same time, the notion of assemblages directs us to consider issues of power (in a Foucauldian [Foucault, 1977] or Butlerian [Butler, 1990] sense) in terms of how some elements of assemblages are negated and others are more durable (Latour, 1990). Given that the durable elements are likely to change over time (Deleuze and Guattari, 1987; Law, 2004), this article addresses three constitutive elements as broad directions to consider – both theoretically and methodologically – rather than claiming them as fixed objects of study. Our proposition is that if we ask questions about: interfaces or content (as momentary stabilisations or representations of socio-technical negotiations [Suchman, 2007]); *alongside* an interrogation of public-facing decision-making processes, which is how we understand community guidelines; *and* machine learning processes as they are also (but differently) momentarily stabilised through recommendation systems, then we can begin to get closer to a sense not only of platforms as assemblages, but also to the durable elements within these assemblages and crucially, what they *do* to the social world. Given this, our article takes as a directive: (1) the content presented to (or concealed from) users through in-platform searches, (2) public-facing community guidelines, which lay out a given platforms' codes of conduct and nod to the political, economic and social considerations of a given social media company and (3) the content that is algorithmically recommended to users as they move through social media.

By evoking the notion of assemblages, we are also of course thinking of the work of Deleuze and Guattari (1987), Law (2004) and DeLanda (2006), particularly for how their use of the term helps us to articulate the intimate connections between and within, for example, the actors and systems that generate communication, design and experience. For Deleuze and Guattari (1987), assemblages seek to explain 'all the voices present within a single voice' (p. 88), and are also 'constantly subject to transformations' (p. 90). Seen here, and relating this work to the scholarship already discussed, are resonances of notions of durability – the 'voices' – and concerns with permutations, multiplicity and processes that Bucher (2018) takes up through her emphasis on performativity (p. 50, following Introna, 2016). These themes are also addressed by DeLanda (2006) when he argues that the performative capacity of assemblages as 'a whole'[1] 'cannot be reduced to those of its parts' precisely because *of* the performative capacity of assemblages: the fact they are 'not an aggregation' of the various components, but 'the actual exercise of their capacities' (p. 11). This suggests, for the purposes of this article, that we need to consider not only the constructions of the various assemblages (our three directives as described above), but also the performative capacity of the assemblages when held together as 'a whole'.[2]

This leads us to our proposition of *sexist* assemblages. Assemblages, as Wiley et al. (2012) note, *do* something, and we argue that only by understanding the particular arrangement of social media content moderation's many elements can we see how they perpetuate rigid gender roles, typically about women; in short, how they perpetuate *sexism*. While we focus on pro-eating disorder (pro-ED) content moderation in this article (a term we define later), we include other examples to argue that the notion of sexist assemblages permeates other platforms and might open up spaces for theoretical and methodological revisions. After discussing social media content moderation and pro-ED identities more generally, we turn to a discussion of the research methods that inspired this article: a cross-platform visual analysis of eating disorder images on Instagram, Pinterest and Tumblr. We then unpack three of the (various) elements of our sexist assemblages – the content presented to and hidden from users through in-platform searches, public-facing community guidelines and recommendation systems – to ask about the implications of such durability within and for assemblages.

## Moderating controversial content on social media

Community-driven Internet spaces have always been moderated in some way, but the growing volume of content uploaded to social media in particular has forced companies to develop more sophisticated moderation techniques. At present, there are two dominant forms of social media content moderation: (1) automated and (2) human. Automated content moderation relies on machine learning techniques which 'consistently maps onto existing data' (Thornham, 2018: 17) in a 'recursive loop' (Day and Lury, 2016: 43): it matches content against known data and databases of 'unwanted' (Roberts, 2017b: n.p.) or flagged content, measuring the distance between points within systems and between certain words or images (Sumpter, 2018: 198). Automated moderation encapsulates the processes of uploading content and the period after: they are both pre-emptive *and* retrospective (Gillespie, 2018). Machine learning moderation compares content with existing data, which means unique content needs to be already normative, or at least 'known' for machine learning moderation to 'see' it as a constitutive element to prompt action, such as deletion. This has a number of implications, but for the purposes of this article also demonstrates why human content moderation is so important for setting the parameters of normativity from which the automated systems can learn and build. When content is flagged, it is often redirected to a human commercial content moderator (CCM) who is given 'seconds' (Roberts, 2017b) to decide if it should stay or go. Content moderation is also outsourced to users who are asked to 'flag' inappropriate content to feed into moderation algorithms (Crawford and Gillespie, 2016). Flagged content is weighted differently within the constitutive elements of algorithms and automated moderation learns from and develops such weighting to create different processes and re-evaluate past outcomes, such as alerts, deletions and restrictions on access to certain content (Suzor, 2016). All of these forms of content moderation are limited for the reasons outlined above, and the automated techniques described above are famously 'imperfect' (Roberts, 2017b: n.p.).

At a macro-level, there are also issues in relation to decisions about what counts as 'problematic' social media content in the first place, not least because of current debates

around social media as on one hand being a safe and supportive space, and on the other, a space that needs safeguarding for vulnerable groups and individuals. As an example, these concerns are reflected in the UK government's new Online Harms White Paper, which lays out plans to develop an independent regulatory body to 'draw up codes of conduct for tech companies', outline their new 'statutory "duty of care" towards their users' and enforce penalties for non-compliance (Goodman, 2019: n.p.). The recent news story about the role Instagram and Pinterest might have played in a teenager's suicide is a case in point (Gerrard and Gillespie, 2019). As Gillespie (2018) argues, moderation of content relating to eating disorders is perhaps 'the hardest to justify' (p. 61), not least because Internet spaces have long been praised (and condemned) for offering non-judgemental communities for those with marginalised or stigmatised identities (among others, Dias, 2003; Turkle, 1996), particularly spaces permitting the use of pseudonyms (Haimson and Hoffmann, 2016; Van der Nagel and Frith, 2015). This contradiction and its accompanying debates have intensified in recent years as tech creators are becoming increasingly aware of the consequences of their designs, while at the same time, as boyd (2015) and Ford (2019) note, being excited by them. Rules about eating disorders were enforced on sites like MySpace, Xanga and Yahoo! but a 2012 *Huffington Post* exposé about Tumblr's 'secret world of teenage "thinspiration"' (Gregoire, 2012) triggered a wave of platform policy alterations. Exactly 2 weeks after the exposé, Tumblr (2012a) released a new policy relating to eating disorder content and Instagram and Pinterest followed suit within the same year. The three platforms said they would draw lines between accounts and posts that 'promote' eating disorders and those aiming to 'build community' (Tumblr, 2012b) or facilitate 'support' (Panzaniro, 2012; Pinterest, 2019) between users. To restrict access to 'problematic' content, the platforms issue public service announcements (PSAs) when users search for certain hashtags or keywords, block or limit the results of searches for other terms, and also remove content and accounts they think breaks the rules (Gerrard, 2018).

Rule-setting is subjective and reflects the biases and worldviews of the rule-setters, and social media's community guidelines are, as Roberts (2019) reminds us, developed 'in the specific and rarefied sociocultural context of educated, economically elite, politically libertarian, and racially monochromatic Silicon Valley, USA' (pp. 93–94). Thus, it is perhaps fair to say that the decision to moderate eating disorder-related content reflects a longer-standing paternalistic desire to 'protect' young women – who are the likeliest gender to experience an eating disorder (among others, ANAD – National Association of Anorexia Nervosa and Associated Disorders, 2019; Beat, 2019) – following a pattern established by traditional media and earlier Internet spaces. The form of sexism we point to in the politics of moderation is also based on a notion of the fragility of *white* women's bodies and the need to protect them. In their analysis of media representations of eating disorders, Saguy and Gruys (2010) found that 'anorexics and bulimics are typically portrayed as young white women or girls, this reinforces cultural images of young white female victims' (p. 231). Eating disorders have long been viewed as a 'White female phenomena' (Root, 1990: 525), an assumption reinforced by misguided research methodologies led by 'stereotypes that only White, middle to upper class girls develop eating disorders' (Root, 1990: 531), and despite evidence to the contrary. But as Lupton (2013) notes, not all women's bodies are seen as fragile nor worthy of protection. In Western

societies, 'It seems that there is something culturally repellent about the fat body, something that calls out to be controlled, contained and punished' (Lupton, 2013: 3), and yet the thin body earns the protection of some of the world's largest corporations. This article attempts to demonstrate how such values can become embedded – or *coded* – into platforms' algorithmic and public-facing methods of control. The platforms' decisions raise a number of questions, but for the purposes of this article they remind us of the wider socio-political arena in which moderation decisions are also being made, and that filters through to how such decisions are variously and unevenly operationalised. We now turn to a discussion of the research methods inspiring this article's content.

## Show and tell: finding images through search results

The findings we present in this article are part of a larger dataset of 975 unique Instagram, Pinterest and Tumblr images. We initially ran searches for 10 keywords using the respective platforms' built-in search engines and used the Digital Methods Initiative's TumblrTool to identify the most common workaround hashtags to give us a set of terms to use in case the root tags were banned.[3] For example, because 'proana' is banned on some platforms, users might search for 'proanaa' or a similar non-banned workaround term (see Chancellor et al., 2016). We originally collected these images to conduct a cross-platform visual analysis, an approach influenced by Ging and Garvey's (2017) finding that images relating to mental health on Instagram are highly aestheticised. Using a clean browser and a new account, we wanted to see what the platforms showed us – a form of platformed *show and tell* – using their search engines to discover new content, precisely as social media users would.[4] But it was the method through which we obtained our data that revealed an algorithmic conflation between posts related to eating disorders and those associated with other feminised phenomena, such as fitness, diet plans, cosmetics and fashion. This resonates with one of the motivations behind Noble's (2018) book, *Algorithms of Oppression*, which she wrote after discovering that Google recommended stereotypical, racist alternatives when she searched for 'Black girls' (as explained in the article's 'Introduction'). Although the flattening of identities Noble (2018) identified resonates with our findings, we want to note that the intersections are different. Noble (2018) found a reproduction of hateful and explicitly sexualised stereotypes of Black women via Google searches, and we identify (as we discuss throughout) a link between white, feminised behaviours and interests with eating disorders via Pinterest recommendations. Although the pattern is similar, the intersections – the 'multiple grounds of identity' including gender *and* race (Crenshaw, 1991: 1245) – are different.

A number of considerations that emerged from the methods and that relate to the concept of sexist assemblages are worth briefly noting here. The first relates to the idea of attempting to momentarily stabilise assemblages through keyword and/or hashtag searches, and the limits and affordances this offers not only in terms of research findings, but also in terms of how these methods can conceptualise platforms 'themselves' (as the sum of these methods). Internet researchers, we suggest, need better methods for capturing the dynamics of social media platforms because *what* we do to understand platforms shapes *how* we understand them. For example, the problem with researching algorithms is that we can only access already-made decisions: it is very difficult to account for

the silences (an issue with assemblage theory itself, and which we return to in our 'Conclusion'). The methodological issues raised here precisely illustrate why this article might be read as both a theoretical and methodological intervention. This means it is important to emphasise that these 'show and tell' methods should not underpin a conceptualisation of the platform, but should rather be understood as a tool through which we might come to understand some of the durable elements of larger processes (see Bucher, 2018). And second, this method highlights the need for dynamic methods that can iterate and change in keeping with algorithms. Keywords and hashtags, for example, are constantly revised and given that any one experience of a platform may be different (see Bucher, 2018), it is particularly important to note an instability not only with the object being studied but also with the methods (Hayles, 2017).

What we found on Pinterest inspired many of the arguments behind this article. For example, when you select a post on Pinterest (on either a mobile app or on the desktop version), you can scroll down to view what Pinterest calls 'more like this': the images you might want to see, based on your browsing habits and other forms of mined social media data (Kennedy, 2016; Sumpter, 2018). The algorithm showed us images related to the root image and also suggested other search terms we might want to explore. When we searched for 'bonespo' – a portmanteau term combining 'bones' and 'inspiration' to denote images which focus on and glorify bones protruding through skin (Cobb, 2017) – Pinterest showed us an image of a seemingly white person's slender legs and small wrist (see Image 1), and suggested we might also like to search for other posts relating to 'grunge', 'hipster' and '90s' fashion (see Image 2).

**Image 1.** The first search result for 'bonespo' on Pinterest.


**Image 2.** Pinterest recommendations following a search for 'bonespo'.

Despite bonespo's clear links to pro-ED discourses, it remained searchable on Pinterest at the time of writing. However, these recommendations alone – and arguably the image itself – are not especially objectionable.[5] They relate to fashion and perhaps classed identities and imply highly stylised gender performances that are consciously intended: something more akin to the notion of self-branding or promotion (Hearn, 2017). What matters here is what prompted the recommendations: a keyword search explicitly related to the promotion of eating disorders. The seemingly mundane process of searching for an image and receiving suggestions for more images users might like reveals a connection between eating disorders, the performatively feminine body and fashion/consumerism. As Dias (2003) notes, and mirrored in the search results discussed above, 'the assumption, evident in most popular notions about eating disorders, is that these women are conforming to dominant notions of femininity' (p. 37). They also conform to dominant notions of white femininity, as grunge subcultures in particular have their roots in 'white youth in the US suburbs' (Huq, 2006: 139). Furthermore, this finding highlights the importance of the 'also liked' algorithm we discuss later in the article.

However, not all pro-ED-related terms are searchable on Pinterest. For example, a search for 'proana' was blocked on Pinterest and prompted the following PSA: 'Are you struggling with an eating disorder? Help is available'. But the failed search for 'proana' also prompted the platform to give us a list of other terms we might want to 'try' (see Image 3).

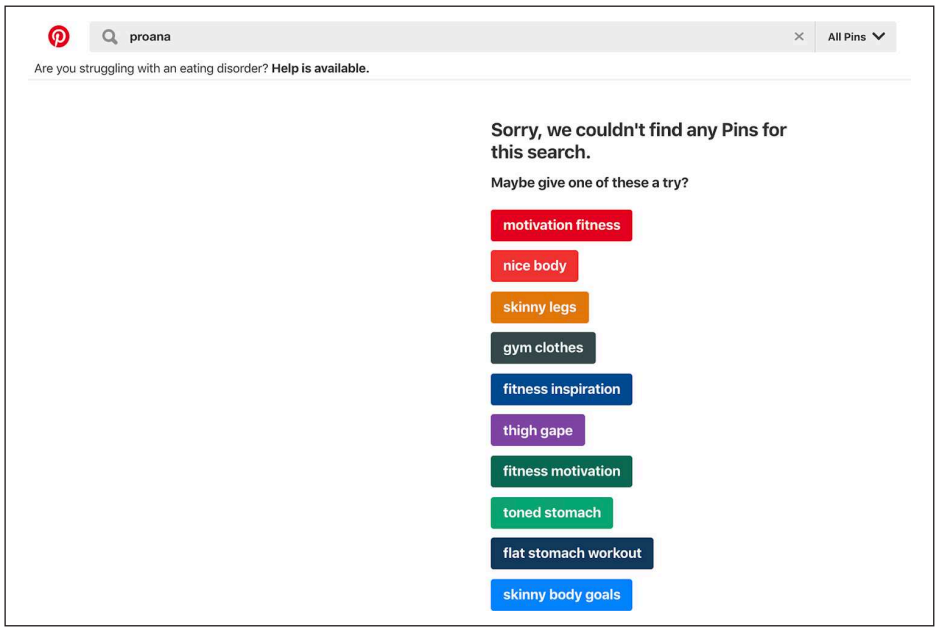



**Image 3.** Recommended search terms to remedy a failed search for 'proana' on Pinterest.

Although the search for 'proana' failed, the platform knew to categorise the term in relation to *bodies* and *body work*, and Pinterest advised us to search for 'nice body', 'skinny legs', 'thigh gape', 'fitness motivation' and 'skinny body goals' instead (Image 3). We then searched for what Chancellor et al. (2016) call a 'workaround' hashtag – #proanaa – and Pinterest gave us the following image (Image 4), followed by further recommendations (Image 5).

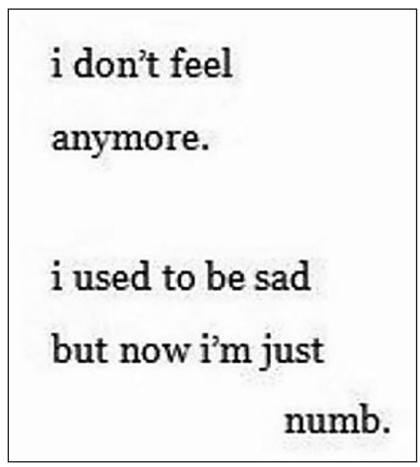



**Image 4.** The first search result for 'proanaa' on Pinterest.

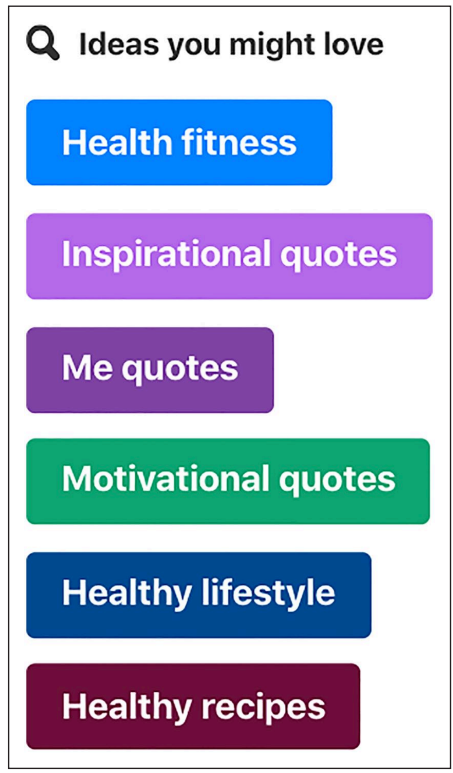



**Image 5.** Pinterest recommendations following a search for 'proanaa'.

At first glance, these recommendations are not explicitly linked to identity markers such as gender, race or age. But when combined – indeed, *assembled* – with both discursive and statistic knowledge of eating disorders indirectly tells us that they are experienced as feminine, white and youthful in multiple ways. Pinterest's recommendations thus add to a broader social imagination and stereotyping of eating disorders not only as a white women's issue but also as a neoliberal, postfeminist preoccupation with body improvement and performances of 'successful' femininity: understanding that femininity is a bodily property, that a woman must possess the *right* body in line with the current hegemonic ideals, that the body must be adequately disciplined and surveilled, and that a woman's 'goals' can be successfully achieved through a makeover (Gill, 2007). The 'sexist' element of the assemblage(s) is bound up in the social and cultural inferences that are both evidenced in the content itself, and through the content in terms of what the search algorithms suggest over time. To re-iterate, this is not to say that these elements are static or will not change but rather that we begin to get a sense of processes and logics that we need to investigate further.

Indeed, to take one of the recommended search terms in Image 3 – 'skinny body goals' – there is both a normativity and mundanity of these associations (given the issues

discussed in the paragraph above) that demonstrate a complicity with the gendering of social phenomena and a misguided alignment of eating disorders with vanity and thinness (Bordo, 2003). At the same time, recommendations are algorithmically generated based on existing data and click-throughs: they represent and perpetuate normativity insofar as they are both an algorithmic outcome of existing activity/behaviour, and they generate and perpetuate ongoing activity/behaviour. In keeping with scholars such as Bucher (2018) and Introna (2016), we are suggesting that recommendations are not transparently gendered (solely) because eating disorders are represented primarily as a female-oriented issue; rather, they are gendered because this conclusion is borne out of existing normative practice and behaviour (see also Neff, 2018).

Returning to DeLanda's (2006) argument that assemblages have performative capacities when held together as 'a whole' (p. 11), our findings also provoke new discussions about the social costs of recommendation systems, particularly if we think about the performative elements of the search algorithms in terms of shaping normativity. One of our main concerns is that search results stabilise, however momentarily, how an eating disorder 'should' be experienced: thin, hyper-feminised, consumerist and by young, white women, not because of the image per se but because of the socio-technical assemblages that have generated it in that moment as an automated response to a query. Our argument here then, is that content, recommendations and searches are all elements of the sexist assemblage, that need to be thought about and investigated together and which also includes algorithmic process and community guidelines and policies. We now turn to a fuller discussion of the latter.

## Community guidelines: the (gendered) rulebooks of social media

The power and politics of social media content moderation not only lie in its processes and outcomes, but also in the decisions about *what* gets moderated and *why* this should happen. This communicative work partly takes place in social media platforms' 'community guidelines', sometimes called 'community standards' or similar. Most, if not all platforms have these public-facing documents and they serve a unique and academically under-addressed purpose: they purport to lay out, in 'deliberately plainspoken language' (Gillespie, 2018: 46), how platforms want their users to behave and what kinds of content they think are and are not acceptable. As suggested earlier, we are interested in the community-facing guidelines as human-machine contextual responses to a perceived human-algorithmic change that are 'caught' momentarily (stabilised) through discourse and that might, in turn, reweigh certain algorithms or change certain processes in visible and less visible ways. But community guidelines are always-already inadequate as a representation of action or policy because the assemblage is in process and iterating beyond that moment. Some 'rules' are more stable than others such as those against supporting terrorism, crime and hate groups, sharing sexual content involving minors, malicious speech and so on, mostly because they verge on or cross the threshold of illegality. But some of platforms' other rules, such as those about eating disorders, are less stable and reflect morality rather than legality.

Community guidelines differ from terms of service and other legal documents because they are intended to be read by users and are written as such. At the very least, community guidelines are spaces in which normativity (as understood by the employees of any given platform) is discussed within a specific temporal and historical context. More than this though, they are spaces where the human rather than the machine comes to the fore, and in juxtaposing the machine learning elements with these discursive human responses, we can see tensions and sutures, priorities and politics. Who responds, when and how is also important not least if we consider, as Gillespie (2018) argues, that the 'voice' of platforms' community guidelines are often consistent with their 'character' (p. 48), which perhaps creates the conditions for them to evoke gendered language.[6] Community guidelines are a crucial component of the assemblage we discuss in this article because they are the spaces where interpretations of values and rules are consciously conveyed. Indeed, while platforms have long emphasised their neutrality (Gillespie, 2010), community guidelines undo some of this careful discursive work by revealing biases, politics and normativities.[7] It is also important to note that community guidelines are also far from static: the guidelines themselves are malleable and constantly being re-shaped and re-purposed; the language changes, the discourses shift. For example, a week after publishing its initial policy, Tumblr issued 'follow-up' guidelines for content related to eating disorders and self-harm and responded to user feedback. One user's comment read:

> It's not a secret that this new rule will target primarily women. Sick women that have finally found a community where they don't feel alone. If you think censoring these websites will lead more women to recovery, consider whether people fought in wars before there was violence on TV. This is shutting down a community where people can talk openly without addressing the (actually evil) blogs that may have caused them to be where they are at. Great job, Tumblr. (Tumblr, 2012b)

The same thing happened when Tumblr announced its ban on adult content – which included images of 'female-presenting nipples' – in late 2018. Following a pattern established by traditional media (among others, see Atwood, 2009; Evans et al., 2010; Gill, 2009), the adult content ban reflected a historic problematisation and over-sexualisation of women's nude bodies. Tumblr then released another post to its Staff Blog clarifying some of the guidelines' details (Tumblr Help Desk, 2018a, 2018b). Community guidelines thus echo other media processes and hint at how those writing the community guidelines see the platform: it is interesting, for example, that Tumblr chose to highlight a comment about sexism in its own follow-up post. These public-facing documents form a core part of sexist assemblages because they (re-)iterate Tumblr's complicity in unequal divisions between acceptable gendered bodies. Evidently, and unlike terms of service, community guidelines are more 'open to outside pressure' (Gillespie, 2018: 70), making them crucial spaces where biases and subjectivities are displayed to users. They offer us insight, we argue, into the politics behind moderation as well as the decisions prompting and responding to machine learning outcomes. If, as Gillespie (2018) notes, 'the full-time employees of most social media platforms are overwhelmingly white, overwhelmingly male, overwhelmingly educated, overwhelmingly liberal or libertarian, and overwhelmingly technological in skill and worldview' (p. 12), we cannot ignore the profound implications this

has on their rules and the broader sexist assemblages we discuss in this article. If we return to our theme of sexist assemblages, we can note the need to also consider issues such as employee demographics, work practices and policies, identity signifiers and labour issues, to name a only few issues at stake. All of these things contribute to people's experiences of platforms, but many are rarely 'seen' or accounted for in the push to only note the productive elements of platforms. This point, then, is a further reminder of the need to consider content moderation as an *assemblage*.

Examples like the above evidence a pervasive platform policing of the female body in particular, not only in the decisions made about the parts of the gendered body that are problematised (protruding bones, female-presenting nipples, etc.), but also, and perhaps even more perniciously, the call within platforms' community guidelines for users to surveil and problematise each other's bodies by flagging content they think glorifies eating disorders. We now turn to a discussion of our final element of the assemblage: social media's algorithmic recommendation systems.

## The 'also liked' algorithm and the (gendered) stakes of recommendation systems

A central way content circulates on social media is through algorithmic recommendation systems, or the 'also liked' algorithm. Such systems are designed to improve user experience, help users to make sense of masses of content, and ultimately retain their participation in – and data-generation on – platforms. But as Sumpter (2018) reminds readers, when faced with a plethora of information, users look at 'fewer options' (p. 107). This is why the also-liked (or 'preferential attachment') algorithm is so powerful: because of how it orders information. Scholars have long argued that we are directed to social media content based on our own *data trails*, which prioritise content in part according to previous activity and purchasing decisions (among others, Cheney-Lippold, 2017; Noble, 2018; Sumpter, 2018). The argument follows that this then leads to a 'filter bubble' whereby our experience online is roughly in keeping with our own socio-political opinions (Pariser, 2011). More recently, however, scholars have questioned the idea of the filter bubble through empirical research that suggests that algorithmic recommendations pay less lip service to variables such as user data and previous data trails and *more* lip service to corporate sponsors and geolocation (Introna, 2016; Noble, 2018: 5). If this is the case, then categories like gender, race and class are being flattened out in keeping with neoliberal and consumerist principles to have economic rather than socio-political resonance (see Thornham, 2018: 127). For this article, these issues demonstrate a need to consider assemblages, content moderation and pro-ED-related content also within economic and consumerist frames.

Sumpter (2018) argues that one of the mathematical formulas applied to social media data is 'principle component analysis' (PCA).[8] PCA works by isolating the strongest correlations in the data and it does this by thematically collating a range of variables into 'cleaner' categories, partly to have fewer categories and therefore stronger correlations (Sumpter, 2018: 29–31). This mathematical sorting prioritises blunt content such as clicks and likes rather than, for example, demographic data or the tone of the post. The

nuances of gender performativity (Butler, 1990) are therefore negated, the tone and style are irrelevant, making the sociocultural and political elements of gender identity flattened and rendered invisible. The 'also-liked' algorithm then bumps up that misreading or simplification of something like gender to grossly exaggerate it as a signifier, and as it increasingly sees this variable, it notes it and gives it more weight. It is the new categories generated through this process that we are concerned with in this article, because what gets generated through recommendation systems are *over-simplified* versions of gender and other identities. To represent only over-simplified versions of gender is, we argue, a form of sexism. This process also plays a crucial role in how phenomena such as eating disorders become linked to certain identities: thin, hyper-feminised, consumerist, youthful and white.

While algorithms do a very good job of appearing to be neutral and wholly driven by user data, they in fact 'represent certain design decisions about how the world is to be ordered' (Bucher, 2018: 67) and as such are selective, partial and constructed (Gitelman and Jackson, 2013; Kitchin, 2014: 14). They are able to 'assign meaningfulness' (Langlois, 2013 in Gillespie, 2014: 167) and are essentially mathematical formulas that come to stand in for gender and other identity markers. Recommendations are an element of sexist assemblages because users receive *more* content which is likely to be gendered in the most simplistic way, as we have previously seen in the images of Pinterest recommendations. Pinterest understands that eating disorders dominantly relate to female bodies, thus in turn users might want to look at other content related to performed femininity such as fashion or cosmetics.

It is particularly interesting to us that recommendation systems do not at first seem to be part of the content moderation process, but this is precisely their power. Recommendation systems and content moderation are not typically discussed together, and this is because they constitute content which is *left over* after moderation has taken place: what other people have 'liked', the accounts they might want to follow, the posts they 'might love'. In other words, recommendations represent the most acceptable content social media has to offer (Gerrard and Gillespie, 2019). Recommendation systems are essentially moderation systems: they are perhaps the most seemingly neutral element of social media and yet the stakes for *how* they categorise content are especially high for how we 'see' gender, along with race, sexual orientation, age, ability and social class. This means recommendation systems are partly responsible for telling users what eating disorders and other social phenomena are, and for reflecting the values of the platform. As Cheney-Lippold (2017) explains, because of the categorisations social media companies make – via decisions about what kinds of content to moderate, how platforms' rules are worded and how they know what to recommend to their users – this means our digital identities are also 'declaration[s] by our data as interpreted by algorithms' (pp. 23–24). We now conclude our article by further considering the social implications of this and other elements of sexist assemblages.

## Concluding thoughts on sexist assemblages and the social

In this article, we have presented three of the many potential elements of what we call *sexist assemblages*: the logics, processes and outcomes of social media content

moderation. We have proposed sexist assemblages as a way of understanding how the human and mechanical elements that make up social media content moderation combine to perpetuate normative gender roles, and to police, perhaps even silence content related to women and their bodies. We have investigated sexist assemblages through three of several potential elements: (1) through the content presented to – or concealed from – users through in-platform searches, and which reflects dominant gendered, racialised and other norms. We argued here that, while problematic, the sexism at play is meant to 'protect' white women; (2) through social media platforms' public-facing community guidelines, which lay out a given platforms' codes of conduct and reveal biases and subjectivity in the decisions about what content to moderate, how best to do so and how these decisions are explained to users. Ultimately, we have argued that pro-ED content moderation reflects longer-standing anxieties around the out-of-control female body (Bordo, 2003) and (3) through the content that is algorithmically recommended to users as they move through platforms. Here, we argued that the seeming neutrality of recommendation systems conceals a powerful process through which eating disorders (and other social phenomena) come to be linked to particular identities. By the time the reader finds this article, the elements of social media content moderation that we identify might have changed, especially given how the elements that make up assemblages are ever-changing, and certain elements are more durable that others (Latour, 1990). But our hope with this article is that in considering the above durable elements and locating them within wider discourses of gender, we can begin to forge a theoretical intervention into how we understand platforms and gender, which has corresponding methodological implications.

We draw this article to a close by making three suggestions for scholars hoping to understand the social in a digital age. First, we underscore the importance of using social media research methods which capture the dynamics of platforms, and to note an instability not only with the object being studied but also with the methods (Hayles, 2017). For example, the problem with researching algorithms, keywords, hashtags and other momentary stabilisations of social media content is that we can only access already-made decisions. It is very difficult to account for the silences, which is indeed an issue with assemblage theory itself. We recognise that one of the main criticisms of assemblage theory is that it only counts or sees the active elements, which creates problems for the unseen or silenced (and which feminist scholarship has long wanted to be attuned to). But assemblage theory helps to direct us to silences; to show us what the most durable elements of an assemblage are; to consider the performative capacity of assemblages when they are held together as 'a whole' (DeLanda, 2006: 11); to tell us what they *do* to the social world.

Second, we call for a recognition and identification of other elements of the assemblage as the ones we propose in this article are not exhaustive. Some might include: social media companies' press releases, public engagement by their representatives, individual decisions made by CCMs, specific decisions made by machine learning systems and users' experiences of gender inequality in moderation decisions. We would suggest that the latter proposal in particular warrants sustained academic interrogation.

Finally, and perhaps most importantly, we urge scholars to continue to engage with the intersectional nature of the assemblages we propose to better understand the link

between content moderation and the social. In addition to highlighting content moderation's parallel protection of whiteness and women, there are places in this article where we note how gender intersects with, for example, sexual orientation and sexuality in Tumblr's adult content ban (Duguay, 2018). Sexist assemblages are thus not only 'sexist' assemblages. We close this article by arguing that the deep embeddedness of sexism within the social – as revealed through the sexist assemblages of the social media platforms discussed here – work to silence some of the most marginal and at-risk social groups, for whom social media promised the strongest community ties.

## Acknowledgements

The authors would like to thank the members of the Social Media Collective, Microsoft Research New England whose thoughtful comments on this research inspired many of the arguments behind this paper. They would also like to thank the anonymous peer reviewers for their generous feedback; Gina Neff for a well-timed comment about the paper's structure; and Elena Maris, Miriam Miller and Joseph Seering for their helpful comments on earlier drafts. Finally, the authors wish to thank Harry Dyer and Zoetanya Sujon for their efforts in organising the *Understanding the Social* conference (January 2019) and for diligently preparing this Special Issue.

## Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

## ORCID iD


Ysabel Gerrard https://orcid.org/0000-0003-1298-9365

## Notes

1. Which is itself a misnomer because the 'wholeness' of an assemblage can only be conceptualised through its relations of exteriority that also change (DeLanda, 2006: 10)
2. While recognising this offers insight but not solutions or answers, and nor can it adequately 'capture' the assemblages-as-processes.
3. At present, a similar co-tag analysis tool does not exist for Pinterest, and the Instagram tool no longer works because of the platform's application programming interface (API) closures. See Reider (2016) for a fuller discussion. We also used the keywords listed by the three platforms when they announced their ban on eating disorder-related content: thinspiration, probulimia, proanorexia (Panzaniro, 2012), anorexia, anorexic, bulimia, bulimic, thinspiration, thinspo, proana, purge, purging and promia (Tumblr, 2012a).
4. This approach is akin to the walkthrough method (Light et al., 2016) but taking algorithmic recommendations as the focus. We maintain this should be a method for Internet researchers – a form of show-and-tell, a way of conducting user research/experiencing platforms like a user would – but it is beyond the scope of this article to develop the method as such.
5. It is beyond the scope of this article to discuss the semiotic consideration of the 'pro'-eating disorder identity, as it is often difficult to know if an image alone, devoid of a caption or text overlay 'promotes' the worsening of eating disorders, but it certainly raises questions about the interpretation of visual imagery both by users and human content moderators.
6. A potential direction for future research might include a discourse analysis of various platforms' community guidelines.

7. As Gillespie (2010) notes, social media companies refer to their products as 'platforms' precisely because of the term's connotations: 'open, neutral, egalitarian and progressive support for activity' (p. 352).
8. The algorithms are black-boxed and they change over time, but Sumpter (2018) makes a highly educated guess based on the basic principles.

## References

ANAD – National Association of Anorexia Nervosa and Associated Disorders (2019) Eating disorder statistics. Available at: https://anad.org/education-and-awareness/about-eating-disorders/eating-disorders-statistics/ (accessed 11 January 2019).

Atwood F (2009) *Mainstreaming Sex: The Sexualization of Western Culture*. London: I.B. Tauris.

Barad K (2009) *Meeting the Universe Halfway*. Durham, NC: Duke University Press.

Beat (2019) Statistics for journalists. Available at: https://www.beateatingdisorders.org.uk/media-centre/eating-disorder-statistics (accessed 11 January 2019).

Bivens R and Haimson OL (2016) Baking gender into social media design: how platforms shape categories for users and advertisers. *Social Media + Society* 2(4): 1–12.

Bordo S (2003) *Unbearable Weight: Feminism, Western Culture, and the Body*. Berkeley, CA: University of California Press.

boyd d (2015) What world are we building? In: Everett C. *Parker lecture*, Washington, DC, 20 October. Available at: http://www.danah.org/papers/talks/2015/ParkerLecture.html (accessed 27 May 2019).

Bucher T (2018) *If. . . Then: Algorithmic Power and Politics*. Oxford: Oxford University Press.

Butler J (1990) *Gender Trouble*. London: Routledge.

Carmi E (2019) The hidden listeners: regulating the line from telephone operators to content moderators. *International Journal of Communication* 13: 440–458.

Chancellor S, Pater JA, Clear T, et al. (2016) #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In: *Proceedings of the 19th ACM conference on computer–supported cooperative work & social computing (CSCW '16)*. Available at: http://www.munmund.net/pubs/cscw16_thyghgapp.pdf (accessed 6 December 2016).

Cheney-Lippold J (2017) *We Are Data: Algorithms and the Making of Our Digital Selves*. New York: New York University Press.

Cobb G (2017) 'This is not pro-ana': denial and disguise in pro-anorexia online spaces. *Fat Studies* 6(2): 189–205.

Crawford K and Gillespie T (2016) What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18(3): 410–428.

Crenshaw K (1989) Demarginalizing the intersection of race and sex: a Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1(8): 139–167.

Crenshaw K (1991) Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Review* 43(6): 1241–1299.

Day SE and Lury C (2016) Biosensing: tracking persons. In: Nafus D (ed.) *Quantified: Biosensing Technologies in Everyday Life*. Cambridge, MA: MIT Press, pp. 43–66.

DeLanda M (2006) *A New Philosophy of Society: Assemblage Theory and Social Complexity*. London: Continuum.

Deleuze G and Guattari F (1987) *A Thousand Plateaus* (trans. B Massumi). Minneapolis, MN: University of Minnesota Press.

Dias K (2003) The ana sanctuary: women's pro-anorexia narratives in cyberspace. *Journal of International Women's Studies* 4(2): 31–45.

Douglas S (2010) *Enlightened Sexism: The Seductive Message That Feminism's Work Is Done*. New York: Times Books.

Drewe N (2016) The hilarious list of hashtags Instagram won't let you search. *The Data Pack*, 10 May. Available at: http://thedatapack.com/banned-instagram-hashtags-update/ (accessed 6 December 2018).

Duguay S (2018) Why Tumblr's ban on adult content is bad for LGBTQ youth. *The Conversation*, 6 December. Available at: http://theconversation.com/why-tumblrs-ban-on-adult-content-is-bad-for-lgbtq-youth-108215 (accessed 7 December 2018).

Duguay S, Burgess J and Suzor N (2018) Queer women's experiences of patchwork governance on Tinder, Instagram, and Vine. *Convergence: The International Journal of Research into New Media Technologies* 26(2): 237–252.

Evans A, Riley S and Shankar A (2010) Technologies of sexiness: theorizing women's engagement in the sexualization of culture. *Feminism and Psychology* 20(1): 114–131.

Ford P (2019) Why I (still) love tech: in defense of a difficult industry. *WIRED*, 14 May. Available at: https://www.wired.com/story/why-we-love-tech-defense-difficult-industry/ (accessed 27 May 2019).

Foucault M (1977) *Discipline and Punish*. New York: Vintage.

Gerrard Y (2018) Beyond the hashtag: circumventing content moderation on social media. *New Media & Society* 20(12): 4492–4511.

Gerrard Y and Gillespie T (2019) When algorithms think you want to die. *WIRED*. Available at: https://www.wired.com/story/when-algorithms-think-you-want-to-die/ (accessed 14 March 2019).

Gill R (2007) Postfeminist media culture: elements of a sensibility. *European Journal of Cultural Studies* 10(2): 147–166.

Gill R (2009) Beyond the 'sexualization of culture' thesis: an intersectional analysis of 'sixpacks', 'midriffs' and 'hot lesbians' in advertising. *Sexualities* 12(2): 137–160.

Gill R (2014) Unspeakable inequalities: post feminism, entrepreneurial subjectivity, and the repudiation of sexism among cultural workers. *Social Politics* 21(4): 509–528.

Gillespie T (2010) The politics of 'platforms'. *New Media & Society* 12(3): 347–364.

Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski P and Foot K (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: MIT Press, pp. 167–194.

Gillespie T (2015) Platforms intervene. *Social Media + Society* 1(1): 1–2.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.

Ging D and Garvey S (2017) 'Written in these scars are stories I can't explain': a content analysis of pro-ana and thinspiration image sharing on Instagram. *New Media & Society* 20(3): 1181–1200.

Gitelman L and Jackson V (2013) Introduction. In: Gitelman L (ed.) *Raw Data Is an Oxymoron*. Cambridge, MA: MIT Press, pp. 1–14.

Goodman E (2019) The Online Harms White Paper: its approach to disinformation, and the challenges of regulation. *LSE Media Policy Project Blog*, 10 April. Available at: https://blogs.lse.ac.uk/mediapolicyproject/2019/04/10/the-online-harms-white-paper-its-approach-to-disinformation-and-the-challenges-of-regulation/ (accessed 16 June 2019).

Gregoire C (2012) THE HUNGER BLOGS: a secret world of teenage 'thinspiration'. *Huffington Post*, 9 February. Available at: http://www.huffingtonpost.co.uk/entry/thinspirationblogs_n_1264459 (accessed 14 June 2017).

Haimson OL and Hoffmann AL (2016) Constructing and enforcing 'authentic' identity online: Facebook, real names, and non-normative identities. *First Monday* 21(6). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/6791/5521

Hayles KN (2017) *Unthought: The Power of the Cognitive Unconscious*. Chicago, IL: Chicago University Press.

Hearn A (2017) Verified: self-presentation, identity management, and selfhood in the age of big data. *Popular Communication* 15(2): 62–77.

Huq R (2006) *Beyond Subculture: Pop, Youth and Identity in a Postcolonial World*. London: Routledge.

Introna LD (2016) Algorithms, governance and governmentality on governing academic writing. *Science, Technology and Human Values* 41(1): 17–49.

Kennedy H (2016) *Post, Mine, Repeat: Social Media Data Mining Becomes Ordinary*. Basingstoke: Palgrave Macmillan.

Kirchner L (2015) When discrimination is baked into algorithms. *The Atlantic*, 6 September. Available at: http://www.theatlantic.com/business/archive/2015/09/discrimination-algorithms-disparate-impact/403969 (accessed 17 May 2019).

Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

Latour B (1990) Technology is society made durable. *The Sociological Review* 38(1): 103–131.

Law J (2004) *After Method: Mess in Social Science Research*. London: Routledge.

Light B, Burgess J and Duguay S (2016) The walkthrough method: an approach to the study of apps. *New Media & Society* 20(3): 881–900.

Lo C (2018) *When all you have is a banhammer: the social and communicative work of volunteer moderators*. Master's Thesis, Massachusetts Institute of Technology (MIT). Available at: https://cmsw.mit.edu/banhammer-social-communicative-work-volunteer-moderators/ (accessed 19 December 2018).

Lupton D (2013) *Fat*. New York: Routledge.

Mackenzie A (2006) *Cutting Code: Software and Sociability*. New York: Peter Lang.

Myers-West S (2018) Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media & Society* 20(11): 4366–4383.

Neff G (2018) *Does AI have gender?* Oxford Internet Institute, 25 June. Available at: https://www.oii.ox.ac.uk/videos/does-ai-have-gender/ (accessed 6 January 2019).

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Panzaniro M (2012) Instagram restricts 'thinspiration' hashtags and bans accounts that glorify self-harm like cutting. *The Next Web*, 21 April. Available at: https://thenextweb.com/apps/2012/04/21/instagram-bans-thinspiration-accounts-and-glorifications-of-self-harm-like-cutting/ (accessed 6 March 2020).

Pariser E (2011) *The Filter Bubble: What the Internet Is Hiding from You*. London: Penguin Books.

Pinterest (2019) Community guidelines. Available at: https://policy.pinterest.com/en-gb/community-guidelines (accessed 11 January 2019).

Reider B (2016) Closing APIs and the public scrutiny of very large online platforms. *The Politics of Systems*, 27 May. Available at: http://thepoliticsofsystems.net/2016/05/closing-apis-and-the-public-scrutiny-of-very-large-online-platforms/ (accessed 4 January 2019).

Roberts ST (2016) Commercial content moderation: digital laborers' dirty work. In: Noble SU and Tynes B (eds) *The Intersectional Internet: Race, Sex, Class and Culture Online*. New York: Peter Lang, pp. 147–159.

Roberts ST (2017a) Content moderation. In: Schintler LA and McNeely CL (eds) *Encyclopedia of Big Data*. New York: Springer, pp. 44–49.

Roberts ST (2017b) Social media's silent filter. *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/ (accessed 19 December 2018).
Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.
Root MPP (1990) Disordered eating in women of color. *Sex Roles* 22(7/8): 525–536.
Saguy AC and Gruys K (2010) Morality and health: news media constructions of overweight and eating disorders. *Social Problems* 57(2): 231–250.
Seering J, Wang T, Yoon J, et al. (2019) Moderator engagement and community development in the age of algorithms. *New Media & Society* 21: 1417–1443.
Squirrell T (2019) Platform dialectics: the relationship between volunteer moderators and end users on Reddit. *New Media & Society* 21: 1910–1927.
Suchman L (2007) *Human-Machine Reconfigurations: Planes and Situated Actions*. Cambridge: Cambridge University Press.
Sumpter D (2018) *Outnumbered: From Facebook and Google to Fake News and Filter-Bubbles – The Algorithms That Control Our Lives*. London: Bloomsbury Press.
Suzor N (2016) How does Instagram censor hashtags? *Medium*, 17 September. Available at: https://digitalsocialcontract.net/how-does-instagram-censor-hashtags-c7f38872d1fd (accessed 29 July 2017).
Suzor N (2019) *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge: Cambridge University Press.
Thornham H (2018) *Gender and Digital Culture: Between Irreconcilability and the Datalogical*. London: Taylor and Francis.
Tumblr (2012a) A new policy against self-harm blogs. *Tumblr Staff Blog*, 23 February. Available at: https://staff.tumblr.com/post/18132624829/self-harm-blogs (accessed 2 June 2017).
Tumblr (2012b) Follow-up: Tumblr's new policy against pro-self-harm blogs. *Tumblr Staff Blog*, 1 March. Available at: https://staff.tumblr.com/post/18563255291/follow-up-tumblrs-new-policyagainst (accessed 2 June 2017).
Tumblr (2018a) Adult content. *Tumblr Help Center*. Available at: https://tumblr.zendesk.com/hc/en-us/articles/231885248-Adult-content (accessed 6 December 2018).
Tumblr (2018b) A better, more positive Tumblr. *Tumblr Staff Blog*, 3 December. Available at: https://staff.tumblr.com/post/180758987165/a-better-more-positive-tumblr (accessed 6 December 2018)
Turkle S (1996) *Life on the Screen: Identity in the Age of the Internet*. London: Weidenfeld and Nicolson.
Van der Nagel E and Frith J (2015) Anonymity, pseudonymity, and the agency of online identity: examining the social practices of r/GoneWild. *First Monday* 20(3). Available at: http://firstmonday.org/article/view/5615/4346
Wajcman J (1991) *Feminism Confronts Technology*. Cambridge: Polity.
Wiley SBC, Becerra TM and Sutko DM (2012) Subjects, networks, assemblages: a materialist approach to the production of social space. In: Packer J and Wiley SBC (eds) *Communication Matters: Materialist Approaches to Media, Mobility, and Networks*. New York: Routledge, pp. 183–195.

## Author biographies

Ysabel Gerrard is a lecturer in Digital Media and Society at the University of Sheffield, UK. Her research has been published in journals like *First Monday* and *New Media and Society*, and her interventions into social media companies' content moderation policies have been discussed in

venues like The Guardian and WIRED. Ysabel is the current Vice Chair of ECREA's Digital Culture and Communication Section, and the Book Reviews Editor for *Convergence: The International Journal of Research into New Media Technologies*.

Helen Thornham is an associate professor in Digital Cultures at the University of Leeds, UK. Her research focuses on gender and technological mediations, data and digital inequalities, embodiment, youth, space, place, and communities. She has led a number of projects investigating practices in digital media that are funded across RCUK. She is author of *Ethnographies of the Videogame: Narrative, Gender and Praxis* (2011) and *Gender and Digital Culture: Between Irreconcilability and the Datalogical* (2018) as well as co-editor of *Renewing Feminisms* (2013) and *Content Cultures* (2014).