The shadow banning controversy: perceived governance and algorithmic folklore

Laura Savolainen
University of Helsinki, Finland

Abstract
In this paper, I approach platform governance through algorithmic folklore, consisting of beliefs and narratives about moderation systems that are passed on informally and can exist in tension with official accounts. More specifically, I analyse user discussions on ‘shadow banning’, a controversial, potentially non-existing form of content moderation on popular social media platforms. I argue that discursive mobilisations of the term can act as a methodological entry point to studying the shifting grounds and emerging logics of algorithmic governance, not necessarily in terms of the actual practices themselves, but in terms of its experiential dimension that, in turn, indicates broader modalities and relationalities of control. Based on my analysis of the user discussions, I argue that the constitutive logics of social media platforms increasingly seem to run counter to the values of good governance, such as clarity and stability of norms, and consistency of enforcement. This is reflected in how users struggle, desperately, to form expectations about system operation and police themselves according to perceived rules, yet are left in a state of dependency and frustration, unable to take hold of their digital futures.

Keywords
algorithmic culture, algorithmic governance, content moderation, content suppression, folklore

Corresponding author:
Laura Savolainen, Centre for Consumer Society Research KTK, Faculty of Social Sciences, University of Helsinki, Snellmaninkatu 10, P.O. Box 16, Helsinki 00014, Finland.
Email: laura.savolainen@helsinki.fi
Introduction

They [YouTube and Google] still shadow ban people, mostly for ideological ‘crimes’. No one ever sees your videos, no one ever reads your comments, you become persona non grata forever. Welcome to the liberal wasteland that is Google.’

‘Shadow banning doesn’t exist. It’s just a myth of the users’.

Shadow banning refers to a controversial and hard to detect type of social media content moderation. The term was reportedly coined by moderators of the Something Awful website and online forum in 2001, to refer to the practice of hiding posts from everyone else except for the poster. Because the problematic user does not perceive the ban, they continue posting to a fictional audience instead of creating a new account. Shadow banning reached wider public awareness in 2018, when conservatives began accusing Twitter of shadow banning ‘prominent Republicans’ by not suggesting them on the platform’s auto-filled drop-down search bar (Stack, 2018). The accusations spawned a debate over the proper definition of shadow banning, as well as triggered a tweet from the then US president Donald Trump, in which he called the moderation technique ‘illegal’: incorrectly, as platforms are guarded by safe harbour laws that grant them broad rights to moderate as they see fit (Gillespie, 2017). Twitter responded to the PR disaster by stating that there had been a glitch, an unanticipated error in the algorithmic system that had limited the search bar visibility of large numbers of users, irrespective of political orientation. By now, shadow banning, or perhaps more accurately, the controversies surrounding it, have become a staple of digital culture. Meanwhile, platforms like Instagram, Twitter and TikTok vehemently deny the existence of the practice.

In this paper, it will be argued that controversies like the one presented above are symptoms of algorithmic platform governance, discussed in this paper as social ordering carried out by social media platforms through the employment of automated means, blending human and machinic agency. If platforms are conceptualised as intermediaries which host and organise user content and interactions, then moderation, the ‘the organised practice of screening and managing user-generated content’ (Roberts, 2019), emerges as their definitional and constitutive task (Gillespie, 2018). Thus, although regulation by platforms spans many domains (like privacy, internet infrastructure, etc.), approaching social media through the lens of emerging moderation practices provides a window into what platforms are and how they govern.

In this article, I approach algorithmic governance from the ground up, through a focus on how users strive to make sense of and address algorithmic power, often failing in their attempts. Based on an analysis of Reddit discussions on shadow banning on prominent social media platforms (TikTok, YouTube and Instagram), I conceptualise shadow banning as algorithmic folklore, consisting of beliefs and narratives about moderation algorithms that are passed on informally and can exist in tension with official accounts. I contend that what is at stake with shadow banning is more than a transient controversy regarding a potentially non-existent form of content moderation. Rather, discursive mobilisations of the term can act as a methodological entry point to studying the shifting grounds and emerging logics of algorithmic governance, not necessarily in terms of the actual practices themselves, but in terms of its experiential dimension that, in turn,
indicates broader modalities and relationalities of control. Therefore, I will conclude the article by contrasting what I have learnt from my data with what is said about and asked of social media platforms as ‘the new governors’ of online spaces (Klonick, 2017), both in the academic literature and by platforms themselves. Most notably, I argue that at the heart of platform governance lies a potentially unresolvable paradox: while platforms are more often taking the stage as legitimate governors, they are also simultaneously proactively developing their moderation practices in a direction that is very clearly not in line with the rule-of-law values of good governance (Suzor, 2018). Before presenting the paper’s theoretical underpinnings, data and methods, and laying out the analysis, a brief overview of what is known about shadow banning on prominent social media platforms is in order.

What is known about shadow banning?

Before laying out my constructivist approach to the phenomenon of shadow banning in more detail, it is useful to consider what is known about the practice on the biggest social media platforms. Academic research on the topic is scant. Are (2021) discusses shadow banning on Instagram through an autoethnographic account. According to her, shadow banning takes place when Instagram ‘dramatically reduces the visibility of posts by hiding them from its Explore page without warning’ (p. 1). Are thus expands the definition of shadow banning to cover not just total invisibility to others, the original referent of the term, but reduced reach as well. A similar diversification of meaning takes place in a recent report by Blunt et al. (2020) on experiences of shadow banning among sex workers and activists on various social media platforms. Approaching the practice through first-person accounts, Blunt et al. sketch many types of shadow ban, including search suggestion ban, action block and a decrease in follower engagement. Both Carolina Are and Blunt et al. discuss shadow bans as a mechanism through which platforms enforce their emerging ‘borderline’ content policies. It is obvious that groups which already exist at the social and cultural margins are vulnerable to being disproportionately affected by such regulations, whether purposefully or due to in-built vulnerabilities in the system.

Contemporary mobilisations of the term shadow banning differ in meaningful ways from the term’s original referent, reflecting trends towards using algorithms and AI in platform governance and content moderation (see Katzenbach, 2021). Indeed, shadow banning had initially nothing to do with algorithms or automated content moderation. What makes the term resonate currently, and sets it apart from other forms of content moderation (like content removals or account suspensions) is that it implies governance that is hidden from the governed.

Besides academic research on user experiences, platform companies have provided some information on the changing contours of their moderation practices, often as a response to accusations of shadow banning. For instance, Instagram has admitted to using a ‘variety of signals, for example, if an account has recently gone against our Community Guidelines, to determine which posts and accounts can be recommended to the community’ (Instagram, 2021a, 2021b). Platforms employ vocabularies of security
and consumer satisfaction, of ‘keeping people safe’ and giving users ‘what they want’ – to justify content suppression. What’s notable about these strategies is their often-anticipatory logic: while content removals take place only after the fact of perceived misconduct, the newer techniques seek to proactively reduce the exposure and impact that potentially problematic content could have, without limitations on visibility.

In a recent tweet thread, Tarleton Gillespie, who is known for his long-standing work on platform politics and content moderation, calls for terminological precision, arguing that people often speak of ‘shadow bans’ when what they are really describing is content suppression. Based on my analysis, I both agree and disagree with his sentiment. Lumping different mechanisms and types of content moderation into the same category, that the use of ‘shadow banning’ by laypeople, journalists, and even researchers often entails, may be counterproductive, because even moderation practices that seem relatively similar may have different policy and free speech implications (Gorwa et al., 2020). Not to mention that in most cases, users can only guess whether, to what extent, and because of what or whom their content was demoted.

However, exactly because of the uncertainty surrounding the moderation practices and algorithmic operations of platforms, I also disagree with Gillespie’s sentiment. If people use the term widely to interpret and share their experiences, then ‘shadow banning’ is an important object of research for academics studying content moderation and algorithmic governance (see also Cotter, 2021). I am not saying that the popularity of the term means researchers should take its existence at face value. Doing so would posit a false equivalency between words and the world; the platforms’ complex, multi-tier moderation mechanisms, and the concepts people come up with to make sense of them. This strategy could even have the adverse effect of hampering our epistemic efforts, as the term easily leads one to see unity, causality and intent where none, necessarily, exists. Rather, in this paper, I approach shadow banning as algorithmic folklore, arguing that by connoting something undisclosed, ‘shady’ and potentially illegitimate, the term functions as a discursive gathering point for the articulation of multiple experiences and beliefs of platform governance, united by the feelings of uncertainty and not knowing.

Content moderation and algorithmic governance on social media

Social media platforms are governed; most importantly, they must function within boundaries set by law. Yet, it is widely accepted that they are also regulators themselves, and that the latter (governance not of but by platforms) matters more in terms of how behaviour and content is structured online (Gillespie, 2017). In this paper, I approach the nature of platforms as governors (Klonick, 2017; Schwarz, 2019; Sinnreich, 2018; Suzor, 2018) (What is their modality of control? How do they relate to their ‘subjects’? Where do they derive their legitimacy from?) through exploring everyday users’ experiences, speculations and strategies regarding algorithmic governance, and more specifically, shadow banning. I see algorithmic governance as socio-technical systems that rank users and content, both assigning them positive relevance or judging them as (potentially) risky or ‘unacceptable’. The promotion and suppression of content are intimately connected: both
operations are guided by rules and values, only some of which are transparent. For example, riskiness and impropriety of content and behaviour are defined by platform policies, and these definitions are only partially laid out in public-facing contracts and documents such as Terms of Service, Community Guidelines and on some platforms, newer Recommendation Guidelines that (vaguely) describe what types of content will not be algorithmically promoted.

In practical terms, platforms’ systems for governing content and users are complex, socio-technical and multi-tier. Sometimes users flag content that is then channelled to human moderators for assessment. Yet, increasingly, algorithms either delete content automatically or flag content for human moderators themselves (Gorwa et al., 2020). Having one’s photos flagged or deleted may hamper one’s future visibility to some extent; for example, Instagram’s algorithms are known to factor in such information (Instagram, 2021a, 2021b). The labour of humans and algorithms is, then, entangled in ways that are increasingly hard to discern, not only in terms of the distribution of work tasks, but also because human agency shapes training data (e.g. what is classified as terrorism, violence, or pornographic nudity), and thus learning algorithms themselves.

Platform policies reflect their position as global, commercial services which need to attract advertisers and create a ‘pleasant’ environment for a culturally and demographically diverse user base, as well as maintain public trust in their ability to self-govern. While earlier writing imagined regulation by platforms as neoliberal, influencing user behaviour through seduction, desire, and the structuring of choice (cf. Beer, 2009; Sampson, 2012), more recent literature has studied platforms as systems of penalisation. Control can be ‘hard’ or ‘soft’ (York and Zuckerman, 2019: 140): ‘a platform’s authority over what can be published online’ on the one hand, and ‘a platform’s authority over what we are likely to see, and what is deprioritised in algorithms that govern a user’s view of posts on the network (the feed)’ on the other.

In the former category, content moderation as post removals has garnered extensive scholarly attention (e.g. Gillespie, 2018; Roberts, 2019). In the latter, Bucher (2012) has argued that platforms promote behavioural codes that align with their business interests, like constant sharing and updating, with an algorithmically enforced ‘threat of invisibility’. In other words, according to Bucher, social media discipline people by pressuring them to work on their (digital) selves according to the norms and valuations that algorithms encode.

The shift in focus towards platform governance as functioning through threat and punishment has offered a new lens to studying social media platforms: in addition to energising users through pleasure and reward, they act as merchant-sovereigns, who dictate the terms of participation, popularity and behaviour online (Schwarz, 2019). Not acting in line may result in exclusion or invisibility, and consequently, the confiscation of social capital that accrues on online platforms (Schwarz, 2019). While state-inspired metaphors of regulation – that treat governance as the capacity to set and enforce rules and provide services (Gorwa, 2019) – illuminate the power and societal role of platforms, they can easily underplay that the ordering of social life that takes place on social media platforms is not just top-down and linear, but in equal parts complex and emergent. Further, users are aware of being subject to ranking, ordering, and punishment. Thus, recent research, explored in the next section, has begun to acknowledge users’
active role in making sense of, acting on, and assigning legitimacy to algorithmic systems, arguing that it must be taken into consideration in order to more fully understand the operation of algorithmic governance.

**Theorising algorithm talk**

Research on people’s empirical engagements with social media algorithms indicates that the complexity and proprietary nature of algorithms does not necessarily foreclose awareness and agency. Novel technologies come across as strange, and inspire efforts to understand them (Star, 1999). However, this is not necessarily thoroughly or technically correct, but may be limited to the extent of being able to tailor them to one’s purposes and embed them in everyday life (Andersen, 2020). While backend data arrangements and the code of social media algorithms may be inaccessible to users, their effects are still felt and perceived (Bucher, 2018). People can observe how their behaviour affects that of the algorithmic system as new inputs, and based on their observations, infer more general rules about its governing logics. Bucher argues that this ability of both humans and algorithms to affect and be affected opens up an agentic space for users to play with, tweak or counteract algorithmic operations.

People combine personal experiences about how platforms work with exogenous information obtained from traditional media, at work or through friends, to develop ‘folk theories’: intuitive, informal beliefs and expectations about system operation (DeVito et al., 2018). As algorithms learn through datafied feedback, popular folk theories (e.g. the idea that YouTube promotes normatively feminine beauty content) shape cultural production and may become self-fulfilling prophecies, even on the system-level (Bishop, 2019). Further, drawing from Hall’s encoding-decoding model of communication, Lomborg and Kapsch (2020) argue that people interpret algorithms not only in terms of their functioning logics, but also normatively; for example, reflecting on whose interests algorithms promote. These normative understandings guide people’s algorithm-related behaviours. While critical users may try to protect themselves from algorithmic seeing (Lomborg and Kapsch, 2020), affirmative users speak of ‘training’ the algorithm: adopting behavioural patterns based on which the algorithm can infer one’s preferences as accurately as possible (Siles et al., 2020). The communicative perspective suggests that users’ interpretations of algorithms, and ways of engaging with them are plural and may diverge from designers’ intentions.

In this paper, I approach shadow banning as algorithmic folklore, in other words, beliefs and stories about algorithmic governance that are passed on informally and can exist in tension with official accounts. Folklore derives from folk, as a ‘group or everyday life’, and lore, as ‘cultural or oral learning and expression’ (Bronner, 2016: 1). It connotes the myths, stories and plays that arise from and refer to the daily life of people and communities (Bronner, 2007). In contrast to canonised works and ideas, folklore is anonymous: passed on from person to person. In this process, it is constantly revised and reinvented in response to changes in context and narrators’ needs and motivations. As a result, its existence is multiple and variable (Bronner, 2016).

At first glance, folklore might seem detached from technology use, the latter often being framed as a rational or technical activity. Yet, in her discussion about narratives
and myths on artificial intelligence, Natale (2019: 713) argues that it is precisely the algorithmic technologies’ opacity that makes them ‘prone to be translated – sometimes even forced – into a variety of narratives that help people make sense of their functioning and presence’. Likewise, Ruckenstein (2021) notes that personal and shared anecdotes about algorithmic operations have become a central source of algorithm-related knowledge, because information about proprietary systems isn’t readily available – and even if it was, understanding how algorithms are built and work would require specialised skills. Algorithmic folklore informs personal algorithm literacies, but it may comprise a kind of ‘fake pedagogy’, leading people to read the digital environment based on ‘wishful, erroneous, or fearful views’ (Ruckenstein, 2021). Thus, in contrast to prior research that associates algorithm talk with enhanced user agency and instrumental action (Bishop, 2019; Bucher, 2018; Cotter, 2019), with the concept of algorithmic folklore, I want to emphasise that engaging in algorithm talk does not necessarily correlate with the ability to effectively address or subvert algorithmic logics. I argue that what is most revealing about shadow banning as algorithmic folklore – in terms of what it tells about contemporary platform governance – are precisely the ways in which it falls short as actionable knowledge and proves itself too uncertain or contestable to function as a basis for self-subjectification, tactical action, or claims-making.

Data and methods

Data collection for the qualitative comment analysis took place in January 2021, with the digital tool 4CAT that enables researchers to filter Reddit comments by keyword. Reddit is a highly popular online forum through which people can create and comment on discussion threads and start or join new subforums. This makes Reddit an effective platform for discovering conversations about topics one wants to study. A downside of Reddit is that information on commenters’ socio-demographic background is limited to what they disclose themselves.

All Reddit comments using the word ‘shadow ban’ were retrieved from six popular Instagram, YouTube and TikTok-related subreddits, resulting in 1676 comments. These were: r/Instagram (1214 posts), r/InstagramMarketing (119 posts), r/socialmedia (63 posts), r/Tiktokhelp (105 posts), r/youtube (147 posts), r/youtubers (28 posts). No period was specified in the search tool, resulting in posts starting from December 2013 until 6 January 2021, the date of query, but the overwhelming majority of comments had been posted during the last 2 years. While this dataset gives a general overview of shadow banning-related experiences and discussions, it by no means captures all potentially interesting data. The moderation practice is undoubtedly discussed on other subreddits as well: ones that are not about social media in general, but focused on more specific activities or topics. Further, keyword searches for alternative ways of spelling shadow ban (e.g. without the space) resulted in over 7000 posts on these six subreddits alone. However, there is no reason to assume that these additional data would have been systematically different due to the variation in spelling.

While discussions on shadow banning did reflect the distinct affordances of the three platforms of interest (TikTok, YouTube, Instagram), the structure of the conversation was strikingly similar, and followed the same pattern across the dataset. Users came to
Reddit to ask for advice regarding whether they were shadow banned on other platforms and what to do about it, as well as to exchange experiences, beliefs, strategies and advice pertaining to social media algorithms. Thus, Reddit discussions enabled me to access both users’ personal accounts of shadow banning, as well as how the practice was socially negotiated. The data were analysed by means of a thematic analysis that is bottom-up and inductive in nature, yet where interpretation is guided by existing literature and domain knowledge. The data were coded with some of the more important categories in terms of frequency of occurrence being evidence (What evidence did people give for being shadow banned? How did they start to suspect it?), expectations and beliefs (What interpretations, expectations and beliefs users expressed regarding the practice?), advice and strategies (What tactics did users suggest, and employ, to avoid or lift perceived shadow bans?), affective expressions relating to algorithms, and finally, consequences of perceived shadow bans on users’ lives.

The empirical analysis is rendered in the form of a temporal sequence, in which users first sense that something is ‘off’ about the performance of their content and start suspecting that their account has been algorithmically tampered with; gather, formulate and exchange speculative knowledge about how algorithms work and what kinds of norms they encode; strive to modify their behaviour according to their interpretations about system operation; and finally, update their algorithm-related beliefs based on the (non-) effectiveness of their strategies. The data were analysed according to a constructivist approach that focuses on how people practically and discursively create the ‘reality’ that is directly inaccessible to them. In other words, people actively construct the world within which they situate themselves as actors. This analytical approach is well suited for my research focus, as I am interested in shadow banning as a socially negotiated vehicle for making sense of and addressing algorithmic governance, rather than as an objectively existing practice (the latter also being difficult access).

**Encountering platform governance**

**Living with numbers**

A recurring expectation expressed by the commenters was that one’s engagement rate and follower counts should follow a linear, positive growth curve. Algorithms were thought to reward engagement and increasing follower numbers with exposure, which was again expected to result in further engagement and followers. Thus, dips, fluctuations and other perceived anomalies in exposure and engagement stood out and gave rise to concerns, as implied by the following quote:

My 3 recent TikToks aren’t reaching FYP [TikTok’s algorithmically curated front page] according to analytics... I am not sure if I am shadow banned, since I can still see my videos under the hashtags. (TikTok)

The quote above, and others like it, illustrate how even users who were uncertain about being shadow banned used the term to frame their situation and story. The situations described rarely fit the original definition of the moderation technique. Typically, the
commenters’ posts were still visible to others, but were perceived to get less visibility and engagement than before. Thus, in and through algorithmic folklore, the uses and meanings of ‘shadow banning’ diversified beyond the term’s formal definition, extending to cover other forms of content moderation as well. Demotion (i.e. reduced reach), for example, was sometimes referred to as ‘soft shadow banning’ in the discussions. Despite its specific original definition, in practice, then, the term shadow banning functioned as a discursive gathering point for the articulation of various experiences of platform governance. What the stories had in common was that they hinted at something happening in the shadows of algorithmic systems – something undisclosed, suspicious and potentially illegitimate.

The data speak of a techno-culture permeated by ubiquitous numbers and characterised by a heightened awareness of them. Quantification is embedded in services through which people construct their identities, work on themselves and pursue their aspirations. Consequently, metrical feedback becomes taken highly personally. In the case of social media metrics, numbers transform from ‘cold’, objective signifiers into measures of individual worth and success that people are emotionally responsive to and may also contest or question in case they contradict other evaluations, such as the evaluation of one’s children as interesting and worthy of attention, or one’s art as meaningful and good:

"Every time I post a video that includes one of my kids, it ends up getting less than 10 views. Am I breaking some kind of rule? How do I get out of the shadow ban?" (TikTok)

"I don’t understand why no one is clicking on my posts. It seems super fishy. I think my art is objectively good, and it’s mostly within a specific fandom . . . I agree that you shouldn’t judge your art based on Instagram engagement, but it’s really hard to not let this stuff get to you."

(Instagram)

The stakes of engagement are even higher as people are drawn to social media not only by an imaginary of participation, but to gain online visibility and popularity (Savolainen, Uitermark and Boy, 2022). Indeed, due to the algorithmic organisation of information, the former may increasingly require the latter. At the level of content, this future-oriented disposition translates into a style of self-presentation known as self-branding. Meanwhile, backstage, it guides everyday users towards behaviours and practices associated with digital marketing (Ruckenstein, 2021). Users keep tabs on how well their content performs, and modify it accordingly (Author removed, 2020). Platforms offer users advanced tools in optimising this feedback loop. TikTok provides an analytics service that displays both account- and post-level information about reception; for example, a graph visualising one’s follower count over time, and a chart breaking down the traffic source for each video. On Instagram, one can switch to a business account to gain access to the platform’s built-in analytics tool, Instagram Insights. It provides detailed information about one’s account and posts: for example, how many unique users a post was displayed to, what percentage of them were followers, and from where the views to the post came from (e.g. hashtags, explore or home feed). Mentions of using these tools to monitor one’s performance were very common in the data.
In users’ stories, one can sense a tension between everyday understandings of personhood and how algorithms see and ‘treat’ individuals (or more precisely, datafied representations of them). Users seemed to think that when someone or something makes a judgement about us, that evaluation should be stable over time. They also seemed to want that evaluation to be holistic, that is, to concern them as whole people. Thus, they were in disbelief when the system suddenly promoted their content less than before; dissatisfied, when some of their posts were afforded less exposure than others. However, algorithms do not assess users as people, but as ‘dividuals’ (Deleuze, 1992). Dividuals are composites of independent features – clusters continuously formed and reformed as the data traces users emit are processed and correlated, real-time, in the cloud. By the same logic, based on its contents and initial reception, the relevance of each post by the same user may be evaluated as different by the system. Treating subjects as dividuals is pragmatic from the point of view of digital capitalism. It enables real-time targeting of content on a highly granular level, and the detection of emergent taste groups and ‘trending’ phenomena and content (as opposed to fixed identifiers, such as gender and age; or predetermined notions of relevance). However, for users this means that they are faced with an experimental environment with inherent dynamics that are rare in everyday life, such as exponential growth. Not only is the environment itself constantly updating, but so is one’s own position and worth within it. In the following comment, the user recognises algorithmic (in)visibility as an outcome of complex and evolving calculations that consider various cues about the poster, the post and the behaviour and performance of other users:

You are not shadow banned. There’s an underlying algorithm that controls who sees your content. . . . You simply have lower engagement than you did before, because something changed in what went into the algorithm for you, so the algorithm is now producing different results. . . . It seems like you have a low number of followers. Something as simple as a single person with a lot of followers who used to engage with your posts but has now stopped could produce this decline. (Instagram)

At the same time, the comment also reiterates a typical feature of shadow banning folklore: its tendency to imagine algorithmic control as performed by one all-powerful algorithm. In practice, platform governance is a multiple and distributed process; different algorithms (such as recommendation and suppression algorithms) may have contradictory objectives, and can even be in conflict with one another in terms of what content should get promoted (Edgerton et al., 2021). In contrast, in users’ stories of shadow banning, the algorithm or the platform company came to personify the complex reality of algorithmic governance:

I’ve had videos that got a lot of views but very few likes, so these probably weren’t good, but if nobody even watches the video, how can they know it’s bad? (emphasis added) (Instagram)

While such personification is technically incorrect or imprecise, it anyhow enabled users to articulate the power imbalance that exists between social media companies and their users; the governors and the governed. Indeed, users often mobilised analogies of
administration, for example pointing out how states or companies in more regulated sectors would be held accountable if they treated citizens and consumers like platforms do:

I’m not sure this could be entirely avoided if YouTube had better policies, but they are certainly due . . . Right now, it is getting away with everything it wants just because it is a big corporation. Can you imagine what would happen if the government decided to strip you away from growth opportunities or shadow ban you? (YouTube)

Reverse chronology is a free democracy. ‘Optimized’ is corporate fascism. (YouTube)

However, as a critique of shadowy or arbitrary use of power, shadow banning folklore has an obvious drawback: because algorithmic predictions of relevance often act as self-fulfilling prophecies, they are difficult to contest. Indeed, claims of suppression could always be dismissed by implying that talent will rise to the top. The algorithms are doing what they are supposed to do. One’s content is not given exposure, because it is simply not interesting enough:

So many people blame low views on shadow ban. Gives them a comfortable excuse to justify their bad content not performing well. (TikTok)

Probing the rules

Users’ stories shed light on the processes through which people arrive at precarious yet somehow actionable knowledge about social media algorithms. In addition to sharing and comparing experiences, it was quite common to report their lay attempts to ‘reverse-engineer’ moderation systems, examining the relationship between input and output to determine ‘how the recipe of the algorithm is composed (how it weights and preferences some criteria) and what it does’ (Kitchin, 2017):

Nobody is entirely sure about the exact definition of ‘shadow-ban’, and it is something out of our own control. . . . Whatever is happening, I’ll keep posting different kinds of content with and without hashtags, and watching the statistics closely to understand what’s going on. (Instagram)

The notion of shadow banning, then, reflects users’ conviction that algorithms organise content according to rules and criteria only some of which are made explicit. They want to follow the rules but are not sure how to do so. Therefore, they adopt an experimental stance towards platforms, probing their contours by way of testing and gaming, and exchanging their hypotheses and results with other users.

Despite the contested nature of shadow banning, there were some stories and beliefs about it that were mentioned more often than others and seemed to be accepted by users. These beliefs are at the core of shadow banning folklore. They could be divided into hashtag-, content- and behaviour-related. Hashtag-related beliefs explained that platforms ban so-called ‘abused’ or misused hashtags, and that any post using such a hashtag will be hidden from others. Content-related beliefs suspected that if a post included curse
words, nudity, alcohol, violence or other ‘borderline’ characteristics, it could be demoted. Most often, users seemed to think that algorithms scan content automatically for such characteristics. According to TikTok users, even bad lighting in a post could result in being deprioritised. They advised each other to delete all borderline content, and make sure that their videos were ‘well-lit’ to regain visibility. Another content-related belief, especially common among Instagram users, explained shadow bans and suppression as the ‘aftermath’ of other moderation practices, that is, having one’s posts flagged or removed – whether legitimately or not. While shadow banning was seen primarily as an algorithmic phenomenon, in these beliefs, it tied in with practices potentially carried out by humans (i.e. other users who flag content, or human moderators who remove it). Some users believed that being shadow banned indicated a kind of algorithmic stigma, entailing intensified, targeted surveillance:

I believe that when your account is in a ‘not optimal standing’ with Instagram (i.e. flagged or shadow banned), Instagram starts checking your stuff closely: images get scanned (for nudity, reposts), meta-data gets read (photos taken from the phone may get exposure, files from photoshop or the internet won’t), and suspicious activity may be blocked (i.e. posting from Chrome) (Instagram)

While moderation is often understood with reference to content, the third category of beliefs, behaviour-related beliefs, explained shadow bans as resulting from users’ interactions with and on the platform. These theories were especially common among Instagram users. As illustrated by the previous quote, according to users, Instagram prefers images taken with one’s smartphone, and actions performed through it (as opposed to the browser version of Instagram). Other behaviour-related theories were that Instagram disapproves the use of third-party applications to manage one’s account, and as the following comment implies, that Instagram seeks to identify artificial, simulated interactions from authentic relationships and engagement:

Yeah, this is a common issue people are experiencing, it’s called a shadow ban. Basically, it’s a new algorithm that Instagram created to fight off spammers, but it captured a lot more accounts than that. Even Instagram can’t lift the ban. (Instagram)

Instagram’s inferred prohibitions on certain behaviours were thought of as an effort by the platform to maintain control over technical affordances and disable people from using the platform in ways that misaligned with Instagram’s idea about what the platform and its user culture should be. However, as the comment makes clear, this doesn’t mean that the platform would have been seen as omnipotent. Algorithmic enforcement of rules was recognised as involving false positives, and as in the case of the previous comment, could even be seen as being out of control, or in control of no one in particular.

**Pleasing the algorithm: role reversal between humans and algorithms**

Users reacted to the speculative knowledge they communally formed and circulated by changing their behaviour. In this section, I study these behavioural modifications in order
to pave the way for asking what users’ attempts, and failures, at aligning themselves with moderation systems say about contemporary platform governance. What became clear in the discussions is that while user behaviour in social media is often analysed with reference to the social dynamics of interaction, people no longer self-present only to other people, but increasingly, also to algorithms:

Lots of innocent people are now suffering from [Instagram going after bots and spammers]. The key is to make Instagram’s algorithm think that you’re just a regular person minding his own business. (Instagram)

As the above quote illustrates, the reason people kept up impressions to algorithms wasn’t because they wanted to engage in ‘borderline’ or unacceptable behaviours without getting punished for it. Rather, Instagram’s algorithm was talked about as hypersensitive, as easily misrecognising sincere behaviours as ‘gaming the system’, human agency as machinic agency, and thus punishing ‘innocent people’. In algorithmic folklore, then, a novel narrative of algorithmic governance began to form, one that serves a counterpoint to the dominant discourse about algorithms as decision-makers. Instead of being rational, objective or ‘rule-based’ (at least in a humanely recognisable way), algorithmic moderation systems were routinely experienced as being prone to error, capricious and tyrannical:

Instagram is weird. Their algorithm is weird, too. Every month we have to draw a probability tree to see what might tick them off and what wouldn’t. It’s really annoying. (Instagram)

Instagram is like an evil puzzle with no correct answer to check with at the end!! Then it all changes again! (Instagram)

Ironically, users had to engage in machine-like behaviours to ascertain algorithms of their humanity and good faith. They exchanged rigid, technical rules for the kind of content to produce, and how many interactions to make per hour or day, in order to grow one’s account without getting mistaken for a bot, and penalised:

Start alternating about 3-9 sets of 30 different hashtags. this way you cover more hashtags to hit top posts – while seeming more genuine than a jarvee.com spam bot (Instagram)

Traditionally, machines have been seen to correct human fallibility (cf. Crawford, 2021). Yet, in the case of social media moderation systems, a role reversal between humans and machines seemingly took place. Faced with the perceived hypersensitivity of algorithms, users sought to align with them by way of compensation, in other words, by becoming more consistent, rational and rule-abiding themselves. The perceived randomness and irrationality of algorithmic decision-making could be seen as rational in the sense that it had disciplinary effects on users. Because of the unpredictability and inscrutability of algorithmic moderation, users had to play it safe. They advised each other to be careful not to step out of line, and produced homogeneous, consistent content as a result. Yet, what is at stake is not merely another story about discipline, only in a less apparent form.
This is because even though users did their best to ‘normalise’ their behaviour in accordance with perceived rules and become ‘docile’ bodies, they routinely failed. These failures forced users to update their views about system operation, and face their own ignorance:

Yeah, it’s their platform, but they’re also a de facto monopoly and it’s not as though any of us have other choices. While I disagree with some of the rules that YouTube has created, I nevertheless abide by the rules they have published. And that’s where things get really frustrating. Over the last year, it has become more and more common for our videos to be [demonetised] for no reason. (YouTube)

Any ideas of what things can cause shadowbans that are usually not known? Literally try to play this by every rule in the book here but somehow get shadowbanned for no reason! (TikTok)

I’m really stretching my brain to try and figure out what the problematic content is. . . . I really wish TikTok would just notify you of whatever the problem is. I can remove music and scenes; I want people to see my videos! I want to follow your mysterious rules, TikTok!!! (TikTok)

The inscrutability of content moderation systems and the inability to address them effectively steered some users towards dystopian conspiracy theories and, relatedly, the politicisation of algorithms as biased against the specific social or political group they identified with. Characteristics of conspiracy theories include a disbelief in the ‘official account’, the assumption of nefarious intent, and the over-interpretation of random occurrences as being part of a broader pattern (Cook et al., 2020):

Social media bosses use Facebook, Google, or YouTube to shadow ban the voices that they don’t like and promote the voices they like. They control the minds of people to make people believe that good politicians are wicked and shady government officials are actually good. (YouTube)

Conspiracy theories challenged the platforms’ official narratives that shadow bans do not exist. They also assumed that social media companies have a sinful agenda, and saw intent in occurrences that might well have been glitches or false positives. At first, conspiracy theories can seem to be silly or untenable, but there is something highly revelatory about the fact that users develop and circulate them. While often seen in a negative light, conspiracy theories have been conceptualised as tools of resistance employed by the oppressed (Turner, 1993), as well as ‘poor person’s cognitive mapping’ (Jameson, 1988: 356), that is, unsophisticated and misguided attempts at imagining and representing the abstract systems underlying and shaping lived experience. These approaches highlight conspiracy theories as perhaps irrational, but still intelligible responses to unequal power relations, crises of representation, and information asymmetries. Indeed, and as demonstrated, if folk theories (understood as a ‘cultural resource from which we can construct strategies for action’ (Nielsen, 2016: 842)) fail to help people operate online, is it a surprise that algorithmic folklore turns towards conspiracy theories instead? While users’ beliefs about algorithms could be incorrect or speculative, the experiences of uncertainty and arbitrariness that they live with are certainly real, and
raise questions about the legitimacy of platform governance as it’s currently being configured.

Discussion and conclusions

In a lot of ways Facebook is more like a government than a traditional company. We have this large community of people, and more than other technology companies we’re really setting policies. – Mark Zuckerberg, CEO of Meta Platforms, the corporation that owns Instagram

By connoting something as undisclosed, ‘shady’ and potentially illegitimate, shadow banning functioned as a discursive gathering point for the articulation of diverse stories and beliefs of platform governance, united by the experience of opacity and not knowing. Through the data, shadow banning emerged as a myth or folktale with which people made meaning of the technological and power relations, and articulated the uncertainties underlying digital experience. Thus, rather than an objectively existing practice, I have approached shadow banning as algorithmic folklore, consisting of stories about algorithmic content moderation that are passed on from person to person, and the existence of which is multiple and variable. This is not to say that users’ stories would not have any grounding in reality; historically, folklore has been falsely associated with error and ignorance (Bronner, 2007: 56–57). Sometimes, the experiences and beliefs users discussed in tandem with shadow banning lined up with what is known about how platforms screen and manage social media content. Algorithmic folklore may, then, provide speculative insights into the changing contours of algorithmic systems (see also Bishop, 2019). Yet, I argue that what is most revealing about shadow banning folklore – in terms of what it tells about contemporary platform politics and relations – are precisely the ways in which it falls short as workable knowledge and proves itself too uncertain or contestable to function as a basis for claims-making. What does it mean for how we should theorise social media platforms as the ‘new governors’ (Klonick, 2017; see also Schwarz, 2019), and what we should expect from them, that despite users’ desperate struggles to form expectations about system operation and police themselves according to perceived rules, they are left in a state of dependency and frustration, unable to take hold of their digital futures?

In Automated Media, Andrejevic (2019) discusses automated decision-making as operational control, as opposed to disciplinary control. Disciplinary power aimed at the subjectification of individuals, and relied on the symbolic efficacy of rules, or signs of surveillance (e.g. security cameras). Operational control does not seek to ‘enforce behavioural norms, but to unleash the full range of activity that will allow patterns to emerge as clearly as possible in order to correlate, predict, and pre-empt’ (p. 40). The fact that users did their best to infer and follow the rules, but were often unable to, points to emerging operational logics of platform governance, rather than disciplinary logics. While operational control may have disciplinary effects, they are not its main aim: rather, platforms strive to control the environment and differently ranked actors’ opportunities to act and become visible within it, rather than their behavioural inclinations themselves. Indeed, by being unclear about certain behaviour norms, platforms can in fact govern more effectively. Given their size and accessibility, potential ‘problematic’ activity and
content (e.g. bots and pornographic nudity) appear to be inevitable and platforms increasingly seek to act on them automatically as they emerge: problematic content is removed or deprioritised before exposure and impact. Disclosing the rules might only result in users trying to ‘game’ them. Thus, it is better to keep users under a cloak of uncertainty with regards to what the rules are, how they are enforced, whether one has broken them, and even whether one is being punished for it, as in content suppression.

Should platforms have a responsibility to inform users about how their content is ‘seen’ and evaluated by algorithms? Or should we approach users’ negative reactions and experiences of algorithmic (down)ranking as an unavoidable trade-off of having algorithms ‘watch’ and ‘listen’ to everything in advance for us, serving as information tasters and testers’ (Andrejevic, 2019: 36)? One drawback of choosing the latter option is that opaque algorithmic curation is effective at generating both sincere and theatrical scepticism. In the latter case, opacity enables politically motivated actors to make neither provable nor easily refutable claims of being discriminated against by social media platforms. Instead of social media serving as fora for democratic participation, their lack of transparency may push discussion towards a performative meta-dispute about biases in the rules of deliberation – which again may hamper deliberative processes themselves by implying that they do not in fact represent all voices. Meanwhile, in the former case, the inscrutability of platform governance (combined with the many scandals surrounding social media companies) leads to the proliferation of conspiratorial beliefs and speculative folk theories about algorithms. While these folk theories are undoubtedly critical, in their dystopian character, they often leave little room for grappling with the real complexities, interdependencies and frictions of algorithmic governance (Ruckenstein, 2021); the collective working out of which is the only way towards a more democratic digital society.

Troublingly though, even if platforms wanted to be more open about the rules that their systems encode and enforce, there are growing technological and epistemic barriers to how much can be disclosed. Judicial processes relying on human moderators are prone to error and overload, and social media platforms have begun automating them with the help of artificial intelligence and predictive algorithms, hoping it would solve said problems. Emerging content suppression practices (for ‘borderline’ content, for example) are often based on the algorithmic processing of multiple signals that are added and removed over time (cf. Instagram, 2021a). At the same time, it has been argued in the contemporary literature in critical algorithms studies that AI and machine learning generate ‘useful but otherwise inaccessible insights’ (Andrejevic, 2019); that the weightings and threshold values based on which algorithms make decisions may not be humanely scrutable or explainable; that because forms of machine learning algorithms learn reactively, by being notified of their mistakes and adjusting themselves accordingly, mistakes and false positives are integral to their training process and functioning (Amoore, 2013: 9); and that these ‘learning algorithms’ ‘constantly revise the rules they enforce’ (Schwarz, 2019: 138). The outcomes of algorithmic governance may be shadowy even to the designers. The organisation of content is a complex, constantly evolving and relational operation, involving, for example, interaction between various input data and, as Facebook’s engineers put it, ‘multiple layers of machine learning models and rankings’ (Lada et al., 2021). While the general logic can be explained, particular decisions may be unintended and unexplainable, even by platform companies themselves. Thus, it seems that
the algorithmic governance of social media platforms can only ever strive for an algorithmically optimised aggregate-level outcome in the long run, not accountability for singular users at present.

To strike a better balance between public values and platforms’ commercial motivations, calls have been made to develop their governance mechanisms in a more transparent and accountable direction. Yet, can we expect social media platforms to become legitimate, consistent regulators, when their very constitutive logics seem increasingly to run counter to ideals such as clarity and stability of norms, and consistency of enforcement? It is indeed ironic that while platforms are more often taking the stage as rational governors – consider Meta Platforms [then Facebook] who in 2020 formed a purportedly independent Oversight Board, whose relationship to Facebook and Instagram is essentially that of a Supreme Court’s to other branches of government – they are also simultaneously proactively developing their moderation practices to a direction that is clearly not in line with the rule-of-law values of consensuality, transparency, equal application, relative stability, and fair enforcement, suggested by the project of digital constitutionalism (Suzor, 2018). Their legitimacy as the ‘new governors’ hinges on the willingness of the public to accept the depoliticised vocabulary of recommendation that connotes serendipity, pleasure and consumer satisfaction. However, while users might affirm recommendation and moderation algorithms as consumers of content, in their capacity as producers, they are likely to be more critical. This is because algorithmic evaluations are often in tension with people’s self-assessments of the worth and appropriateness of their content. However, as people try to make sense of their content’s poor performance, they sometimes see malintent where none exists. This may lead them further away from the real ethical and political issue: not shadow banning, but shadowy governance by complex, distributed systems that rank and order social life at a global scale, yet aren’t fully in control of anyone in particular, and are difficult to meaningfully understand and make sense of.

Acknowledgements

Special thanks to the organisers of The Digital Methods Initiative’s Winter School 2021 on ‘Bias in content recommendation and moderation’ and my teammates Caitlin Harzem, Mea Lakso, Karlygash Nurdilda and Clemence Degrandi, who sparked my interest in the topic of shadow banning and with whom we collected the online forum data. I am also grateful for Assoc. Prof. Minna Ruckenstein (University of Helsinki) for insightful discussions that helped me in developing the analysis.

Funding

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

ORCID iD

Laura Savolainen https://orcid.org/0000-0002-5979-1607

Notes

1. The quotes are excerpts from the data I analysed for this paper.
2. As noted, content moderation is a multi-tier system involving human and machine agencies.
However, as my empirical material shows, shadow banning is made sense of primarily as an algorithmic phenomenon. Therefore, it comprises algorithm-related – rather than some other type of – social media folklore.

References

Amoore L (2013) *The Politics of Possibility: Risk and Security Beyond Probability*. Durham, NC: Duke University Press.

Andersen J (2020) Understanding and interpreting algorithms: toward a hermeneutics of algorithms. *Media, Culture & Society* 42(7–8): 1479–149.

Andrejevic M (2019) *Automated Media*. New York, NY: Routledge.

Are C (2021) The Shadowban cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*. Epub ahead of print 19 May 2021. DOI: 10.1080/14680777.2021.1928259

Beer D (2009) Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society* 11(6): 985–1002.

Bishop S (2019) Managing visibility on YouTube through algorithmic gossip. *New Media & Society* 21(11–12): 2589–2606.

Blunt D, Wolf A, Coombes E, et al. (2020) Posting into the Void: studying the impact of shadow-banning on sex workers and activists. Report, Hacking/Hustling. Available at: https://hacking-hustling.org/posting-into-the-void-content-moderation/

Bronner SJ (2007) *The Meaning of Folklore*. Logan: Utah State University Press.

Bronner SJ (2016) *Folklore: The Basics*. London and New York, NY: Routledge.

Bucher T (2012) Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society* 14(7): 1164–1180.

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

Cook J, van der Linden S, Lewandowsky S, et al. (2020) Coronavirus, ‘Plandemic’ and the seven traits of conspiratorial thinking. *The Conversation*, 15 May. Available at: https://theconversation.com/coronavirus-plandemic-and-the-seven-traits-of-conspiratorial-thinking-138483 (accessed 1 November 2021).

Cotter K (2021) “Shadowbanning is not a thing”: black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*: 1–18. Epub ahead of print 28 October 2021. DOI: 10.1080/1369118X.2021.1994624.

Cotter K (2019) Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society* 21(4): 895–913.

Crawford K (2021) *The Atlas of AI*. New Haven, CT: Yale University Press.

Deleuze G (1992) Postscript on the societies of control. *October* 59: 3–7.

DeVito MA, Birnholtz J, Hancock JT, et al. (2018) How people form folk theories of social media feeds and what it means for how we study self-presentation. In: *CHI ’18: Proceedings of the 2018 CHI conference on human factors in computing systems*, Montreal, Canada, 21–26 April, pp. 1–12.

Edgerton A, McLaughlin D and Campbell M (2021) How Facebook algorithms can fight over your feed. *Bloomberg*, 25 October. Available at: https://www.bloomberg.com/news/articles/2021-10-25/how-facebook-algorithms-can-fight-over-your-feed-quicktake (accessed 1 November 2021).

Gillespie T (2017) Governance of and by platforms. In: Burgess J, Poell T and Marwick A (eds) *SAGE Handbook of Social Media*. London: SAGE Publications, pp.254–278.

Gillespie T (2018) *Custodians of the Internet*. New Haven, CT: Yale University Press.

Gorwa R (2019) What is platform governance? *Information Communication & Society* 22(6): 854–871.
Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7(1): 1–10. DOI: 10.1177/2053951719897945

Instagram (2021a) Why are certain posts on Instagram not appearing in Explore and Hashtag pages? Available at: https://www.facebook.com/help/instagram/613868662393739 (accessed 1 November 2021).

Instagram (2021b) Shedding more light on how Instagram works. Available at: https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works

Jameson F (1988) Cognitive mapping. In: Nelson C and Grossberg L (eds) *Marxism and the Interpretation of Culture*. Urbana, IL: University of Illinois Press, pp. 347–358.

Katzenbach C (2021) “AI will fix this” – The technical, discursive, and political turn to AI in governing communication. *Big Data & Society* 8(2): 1–8. DOI: 10.1177/20539517211046182.

Kitchin R (2017) Thinking critically about and researching algorithms. *Information Communication & Society* 20(1): 14–29.

Klonick K (2017) The new governors: The people, rules, and processes governing online speech. *Harvard Law Review* 131: 1598–1670.

Lada A, Wang M and Yan T (2021) How does News Feed predict what you want to see? Personalized ranking with machine learning. *Tech@Facebook*, January 26. Available at: https://tech.fb.com/news-feed-ranking/ (accessed 1 November 2021).

Lomborg S and Kapsch P (2020) Decoding algorithms. *Media, Culture & Society* 42(5): 745–761.

Natale S (2019) If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. *New Media & Society* 21(3): 712–728.

Nielsen RK (2016) Folk theories of journalism: The many faces of a local newspaper. *Journalism Studies* 17(7): 840–848.

Roberts ST (2019) *Behind the Screen*. New Haven, CT: Yale University Press.

Ruckenstein M (2021) *The feel of algorithms: Data power, emotions, and the existential threat of the unknown*. [Unpublished manuscript].

Sampson TD (2012) *Virality: Contagion Theory in the Age of Networks*. Minneapolis, MN: University of Minnesota Press.

Savolainen L, Uitermark J and Boy J (2022) Filtering feminisms: Emergent feminist visibilities on Instagram. *New Media & Society* 24(3): 557–559.

Schwarz O (2019) Facebook rules: Structures of governance in digital capitalism and the control of generalized social capital. *Theory Culture & Society* 36(4): 117–141.

Siles I, Segura-Castillo A, Solís R, et al. (2020) Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South. *Big Data & Society* 7(1): 1–15. DOI: 10.1177/2053951720923377.

Sinnreich A (2018) Four crises in algorithmic governance. *Annual Review of Law and Ethics* 26: 181–190.

Stack L (2018) What is a ‘Shadow Ban’, and is Twitter doing it to republican accounts. *New York Times*, 26 July. Available at: https://www.nytimes.com/2018/07/26/us/politics/twitter-shadowbanning.html

Star SL (1999) The ethnography of infrastructure. *American Behavioral Scientist* 43(3): 377–391.

Suzor N (2018) Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society* 4(3): 1–11. DOI: 10.1177/2056305118787812

Turner PA (1993) *I Heard it Through the Grapevine: Rumor in African-American Culture*. Oakland, CA: University of California Press.

York JC and Zuckerman E (2019) Moderating the public sphere. In: Jorgensen RF (ed.) *Human Rights in the Age of Platforms*. London: MIT Press, pp.137–162.