Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/CLSR

Computer Law & Security Review

Comment

An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation

Paddy Leerssen

University of Amsterdam, Institute for Information Law, Netherlands

ARTICLE INFO

Keywords:
Platforms
Content moderation
Content curation
Shadow banning
Transparency

ABSTRACT

This paper offers a legal perspective on the phenomenon of shadow banning: content moderation sanctions which are undetectable to those affected. Drawing on recent social science research, it connects current concerns about shadow banning to novel visibility management techniques in content moderation, such as delisting and demotion. Conventional moderation techniques such as outright content removal or account suspension can be observed by those affected, but these new visibility often cannot. This lends newfound significance to the legal question of moderation transparency rights. The EU Digital Services Act (DSA) is analysed in this light, as the first major legislation to regulate transparency of visibility remedies. In effect, its due process framework prohibits shadow banning with only limited exceptions. In doing so, the DSA surfaces tensions between two competing models for content moderation: as rule-bound administration or as adversarial security conflict. I discuss possible interpretations and trade-offs for this regime, and then turn to a more fundamental problem: how to define visibility reduction as a category of content moderation actions. The concept of visibility reduction or 'demotions' is central to both the shadow banning imaginary and to the DSA's safeguards, but its meaning is far from straightforward. Responding to claims that demotion is entirely relative, and therefore not actionable as a category of content moderation sanctions, I show how visibility reduction can still be regulated when defined as ex post adjustments to engagement-based relevance scores. Still, regulating demotion in this way will not cover all exercises of ranking power, since it manifests not only in individual cases of moderation but also through structural acts of content curation; not just by reducing visibility, but by producing visibility.

1. Introduction

Content moderation knows no shortage of scandals. From Twitter suspending Donald Trump to YouTube banning Alex Jones, the public record is rife with controversy, debate, and backlash. And yet, speculation abounds that many more cases may be hidden from view. "Shadow banning", as it has come to be known, alleges that platforms intervene in subtler ways, not by suspending users outright but by secretly demoting them in their recommender systems.

E-mail address: p.j.leerssen@uva.nl

https://doi.org/10.1016/j.clsr.2023.105790

Accusations of shadow banning elicit conflicting responses. For some, it is mere paranoia, stemming from misunderstandings about the ways platforms curate content. Others agree that shadow banning exists, but disagree as to its merits. Is it devious and undemocratic subterfuge, repugnant to fundamental rights and due process? A harsh but necessary defence against social media's most persistent bad actors? Or simply an unintended by-product of new visibility management techniques in content moderation? These questions have become all the more pressing as the EU moves to regulate due process and transparency for content moderation in its new Digital Services Act (DSA). This law attempts to settle the shadow banning question: when, if at all, should content moderation decisions be allowed to remain secret?

This paper offers a legal perspective on the shadow banning phenomenon. Drawing on recent research in the social sciences and humanities, it starts by analysing shadow banning in terms of its terminology, techniques, and policy drivers. Then it examines how the DSA regulates shadow banning through its new due process framework for content moderation, and how this legislation attempts to balance conflicting interests in transparency and secrecy. The final section critiques the concept of visibility reduction or 'demotion', which is central to both the shadow banning imaginary and the DSA's response to it. I review the challenges in defining and enforcing demotion as a legal category, and its limitations in checking the more structural dimensions of ranking power as content curation.

2. 'Shadow banning' as a function of visibility remedies

This section introduces the shadow banning phenomenon. It discusses the concept of shadow banning, the content moderation techniques involved, and the policy considerations driving this phenomenon. My core claim is that shadow banning refers primarily to output-based forms of opacity: the *what* of moderation, not the *why*. Shadow banning concerns therefore revolve mainly around novel visibility reduction techniques, which are output-opaque, rather than conventional content moderation techniques such as account suspension and content removal, which are generally self-evident to those affected. In this way, shadow banning discourse reflects heightened transparency concerns associated with the move in content moderation towards visibility reduction.

2.1. *Definitions: what is shadow banning?*

The term "shadow banning" is colloquial in origin and its usage has changed over time. Originally, the term referred to a deceptive type of account suspension on web forums: a shadow banned user would be give the impression that they were still able to post, whereas in fact their content was no longer visible to any other users (Radsch, 2021). Some sources continue to use the term in this way (including, as we will see, the DSA itself). But in most recent usage, shadow banning usually refers to alternative remedies, especially visibility remedies such as delisting and downranking (Cotter, 2021). These remedies do not cut off access to content entirely, but instead make this content less visible through discovery features such as search and recommendation.

It is in this new, broader form, that talk of "shadow banning" has become prevalent in popular and academic discourses. In 2018, US president Donald Trump accused social media firms without evidence of "shadow banning" conservative viewpoints (Radsch, 2021). Elon Musk, during his takeover of Twitter, tweeted ominous imagery of a shadowy cabal he described as the "Twitter Shadow Ban Council" (Nicholas, 2022). On the other end of the political spectrum, shadow banning allegations have also been raised by marginalised groups including online sex workers and LGBT+ users, as well as by climate activists (Are, 2021; Griffin, 2022; Lulamae, 2022; Nicolas, 2022). Concurrently, social scientists have also started inquiries into shadow banning (Myers West, 2018; Cotter, 2021; Le Merrer et al., 2021; Jaidka et al., 2022; Horten, 2021). Some have tried to detect shadow banning using computational methods, while others have investigated user experiences and perceptions. These studies tend to define shadow banning broadly, for instance, as "a wide range of techniques that artificially limit the visibility of targeted users or user posts" (Le Merrer et al., 2021).

What seems to unite the previous and present meanings of shadow banning, is a particular form of secrecy. Whether as account suspension or as visibility restriction, shadow banning has always referred to content moderation sanctions which the affected user is unable to detect. Shadow banning discourse therefore articulates a distinct type of transparency-based critique; whereas much criticism addresses the uncertain *grounds* for content moderation, and asks *why* certain items have been actioned and not others, shadow banning speaks to a prior question: *what* items have been actioned in the first place? (c.f. Gorwa et al., 2020; Suzor et al., 2019) In algorithmic terms, shadow banning speaks to an opacity of content moderation's outputs, rather than logics or inputs.

Shadow banning therefore raises distinct normative concerns; compared to unexplained sanctions, secret sanctions are even more difficult to hold to account or resist. From a legal perspective, unexplained sanctions are problematic because they thwart opportunities for reasoned contestation, appeal, and hence due process (Waldron, 2016). But secret sanctions go further still, precluding practically all possibilities for individual and collective resistance – whether through legal remedies or through social or political pushback (q.v. Cobbe, 2021). This makes shadow banning an especially powerful and controversial form of secrecy.

A question I will bracket for now, is how to define content moderation sanctions such as visibility reductions. In many shadow banning disputes, I will argue further below, the true disagreement may not be an empirical – has the item been moderated or not? - but rather conceptual – what does it mean for an item to be moderated? In the context of algorithmic ranking and visibility management, many practices tread an unclear line between content moderation, as a process of content classification and enforcement through sanctions, versus content curation, as the process through which platforms select for relevance and "filter abundance into a collection of manageable size" (Thorson and Wells, 2016). I return to this problem below.

2.2. *Techniques: how do platforms shadow ban?*

Shadow banning is at once a matter of policy and of design. As a matter of policy, shadow banning is per definition a sanction which is not disclosed to the affected user. But as a matter of design, some sanctions can still be observed by users even when they are not disclosed, and even despite deliberate efforts to conceal them. Content moderation leaves "traces" (Gillespie, 2022), and some remedies leave clearer traces than others. Shadow banning occurs, therefore, when a traceless remedy is not disclosed. Below I will argue that the conventional methods of takedown and account suspension are relatively self-evident even when platforms try to conceal them, and therefore do not afford effective shadow banning. Visibility remedies, by contrast, leave little or no trace and result in shadow bans by default. This makes the policy question of moderation transparency rights all the more salient for these novel techniques.

Content takedown and account suspension are self-evident because they cut off engagement by all other users (views, likes, comments, and so forth). Platforms may try to conceal this fact by presenting an alternative reality to the affected user, giving them the false impression that their content is still online whereas in fact nobody else can see it. But these methods are unlikely to mislead users for long, since they cause all engagement to grind to a halt. All but the least popular uploaders, therefore, are likely to notice that something is amiss. And since these takedowns and suspensions cut off all engagement, any suspicions are relatively straightforward to test, for instance by logging off, switching to a different account, or asking a friend to check for access. For these reasons, takedowns and suspensions do not afford enduring secrecy, even when platforms try to conceal them.

Visibility remedies, by contrast, tend to be subtler. Their precise effects vary, since visibility remedies can take various forms (Goldman, 2021; Gillespie, 2022). Platforms can remove content entirely from a given feature ('delisting'), reduce its relative prominence within that feature ('demotion'), or impose some other restriction such as a disclaimer or warning screen. These modalities can in turn apply to different recommendation (sub)systems. For instance, Twitter's arsenal of visibility remedies includes search delisting; search suggestion delisting; and "reply deboosts", which demote the target's replies to the bottom of the page and hide it behind a "show more replies" prompt (Jaidka et al, 2022). In theory, visibility restrictions can also be personalised towards specific audiences, and hide an item from certain cohorts or demographics but not from others. Through these and other features, platforms conduct a complex "management of visibilities" that steers and nudges audiences in more or less subtle ways (Flyverbom, 2019).

The problem with observing visibility remedies is, in essence, that visibility on platforms fluctuates constantly and on a personalised basis. Content visibility is governed by complex recommender and search systems, which operate through recursive interactions between user behaviour and machine-learning optimisation algorithms (Leerssen, 2020). In this dynamic, volatile process of content curation, visibility restrictions are simply one factor out of very many, and their impact on overall outcomes may be difficult or even impossible to discern (Jaidka et al., 2022; Le Merrer et al., 2021; Horten, 2021). And since visibility outcomes are personalised to individual users, even observing these outcomes at a systemic level is challenging (Bodo et al., 2017).

Visibility restrictions are most noticeable when they cause steep drops in an item's traffic (Cotter, 2021). But even this is not conclusive evidence. The same drops can also be attributed to user-related changes such as weakening audience engagement or intensified competition from rival uploaders (Gillespie, 2022). The cause could also be a structural change to the platform's ranking methods, rather than an individually targeted sanction. These competing explanations are difficult to rule out since most platforms do not disclose detailed recommendation and engagement data to their uploaders. Most platforms offer little more than aggregate counts of views and engagements, and not the types of recommendation and analytic data that would allow users to observe shadow banning's effects, and distinguish them from more routine operations (Section 4.2 below).

Allegations of shadow banning in specific cases tend to be ignored by platforms, or responded to only partially, even though they admit at a policy level to using visibility restrictions.[1] For Kelley Cotter (2021), this strategy amounts to a form of "gaslighting", which maintains the platform's "epistemic authority" over shadow banning allegations while delegitimising valid concerns as paranoia or conspiracy theory. The platform's epistemic authority over visibility remedies is not absolute, however. Academics and other experts have been able to demonstrate undisclosed demotions. By collecting ranking data at scale, with the help of bots or user participants, one can detect especially drastic and targeted changes to recommendation trends, which permit few other explanations than a targeted restriction (Jaidka et al., 2021; Le Merrer et al., 2021; Horten, 2021). But these sophisticated measures are out of reach for the vast majority of users. And in theory, platforms could minimise the risk of detection by designing their downranking measures adversarially, for instance by downranking items gradually over time rather than instantaneously (Pasquale, 2015, p. 285). By the same logic it follows that the most restrictive delisting measures are relatively more easy to detect than subtler forms of demotion, since their effects are more pronounced. Researchers have been able to create tools that test for delisting automatically, such as Shadowban.eu and Whosban.eu (Le Merrer et al., 2021). These tools can instantly test whether specific accounts have been delisted from Twitter's search and autosuggest features by querying relevant

[1] Platform denials are often based on restrictive (and perhaps misleading) conceptions of shadow banning. One official statement post by Twitter (2018) denied shadow banning, defined as: "deliberately making someone's content *undiscoverable to everyone except the person who posted* it, unbeknownst to the original poster" (emphasis mine). Their statement, though nominally denying shadow bans, fails to clarify whether non-takedown remedies such as downranking are being applied without notice. Kelley Cotter (2021) observes a similar strategy in Instagram's communications: "while Instagram's statements avoid obvious falsehoods, they omit important clarifying information, for example a clear and consistent definition of shadowbanning".

phrases, but for subtler demotions such tests would be more challenging. Ironically, then, for their victims and for the public at large, visibility remedies are often invisible.

An additional category of opaque moderation techniques is demonetisation, which renders items ineligible for advertisement revenue-sharing programs (i.e. 'monetisation'). In a study of YouTube's policies, Robyn Caplan and Tarleton Gillespie (2020) note that demonetisation can also be difficult for users to observe. Much like visibility restrictions, the problem with observing demonetisation stems from a combination of volatile engagement patterns and a lack of granular data access. Since YouTube's disbursement statements did not break down revenue for individual videos, users were usually unable to discern whether any of their videos might have been demonetised—let alone establish which videos in particular had been actioned. In 2018, YouTube changed course and started disclosing monetisation status on a per-video basis. This newfound transparency prompted vigorous criticism and resistance from users, who saw inconsistency and discrimination in YouTube's decisions (Caplan and Gillespie, 2020). Some of these users sought to hold YouTube accountable through public criticism and awareness raising, whilst others resisted the policy by switching to other platforms or other revenue models (e.g. direct donations). This episode speaks to the importance of notice policies for unobservable remedies such as demonetisation, delisting and demotion. With notice, they are resisted. Without notice, they result in shadow bans.

2.3. *Policies: why do platforms shadow ban?*

In light of the above, it should be clear why shadow banning concerns revolve primarily around visibility remedies, and, to a lesser extent, demonetisation. The basic problem with these remedies is that, unless notified, users struggle to ascertain whether or not they have been sanctioned. Explaining this phenomenon therefore entails two discrete questions: Why do platforms deploy visibility restrictions? And why do they refrain from notifying them?

To start with the first question: platforms have recently started intensifying their use of visibility reductions as a supplement to conventional moderation strategies. In particular, visibility remedies are used to manage new controversies which often fall short of violating established laws, such as disinformation, hate speech, and 'clickbait'. To justify intervention on such issues, and deflect accusations of censorship, platforms and policymakers alike have touted visibility reduction as a less restrictive alternative to removal. "We're not arguing for censorship, we're arguing just take it off the page, put it somewhere else.", Google CEO Eric Schmidt has claimed (Wisner, 2017). For Facebook CEO Mark Zuckerberg, visibility reduction would help platforms to manage disinformation without becoming "arbiters of truth" (Swisher, 2018). Rather than remove disinformation, Zuckerberg argued, "we feel like our responsibility is to prevent hoaxes from going viral and being widely distributed" (Swisher, 2018). This turn to visibility remedies forms part of a broader reframing of platform culpability from *publication* to *amplification*, i.e. the granting of excessive visibility (Keller, 2021; Miller, 2021). Its slogan: Renee DiResta's widely-cited adage that "free speech is not free reach" (Diresta, 2018).

What is missing from this account is the problem of transparency. Given that visibility remedies are less noticeable than conventional sanctions, and therefore result in shadow banning, they are arguably *more* restrictive for users, not less (Horten, 2021). The perverse result is a type of reverse proportionality: the most sensitive edge-cases end up being governed by the least transparent means. Instead of a Ministry of Truth, we get a secret police.

Why, then, are most visibility remedies not disclosed to users? One factor may be the cost and complexity of disclosure; Gillespie (2022) notes that visibility restrictions are more complex than other sanctions, and not as amenable to meaningful disclosure. The sheer novelty of these techniques might also explain, in part, why disclosure practices have not yet caught up. But besides mere cost and novelty, platforms may also have more deliberate reasons to maintain secrecy. For instance, Monica Horten (2021) sees shadow banning measures as a strategy grounded in the adversarial logics of computer security. From the moderator's perspective, secret sanctions can be a convenient way of mitigating resistance and adaptation from persistent rules violators such as commercial spammers. If notified, these users might for instance respond by creating new accounts, or by attempting through trial and error to reverse engineer the platform's classification methods so as to 'game the system' and evade detection altogether (Cotter, 2019). Still, what counts as legitimate compliance and what amounts to illegitimate 'gaming' is determined by the platform itself and in practice often deeply ambiguous (Cotter, 2019; Poell et al., 2022). That ambiguity may be problematic from a legal due process perspective, which insists on rule-bound and foreseeable sanctions. But from the moderator's perspective, notifying these sanctions and clarifying ambiguities may only serves to lessen one's control over the service. For the platform, shadow banning may then be a feature, not a bug. From a public interest perspective, tensions emerge between ideals of due process and accountability on the one hand, and the need for effective content moderation on the other hand.

Although defences of shadow banning are often cast in the technocratic language of security and circumvention, there are also political and reputational interests at stake. Platforms are unlikely to admit it, but opaque visibility remedies can be a strategy to avoid public accountability (Gillespie, 2022). Platforms may well claim to embrace transparency and accountability, but their track records show the opposite; many important transparency reforms are only made under public pressure or legal obligation (Zalnierute, 2021). It is clear that platforms see content moderation as a source of reputational risk, and secrecy as a means to mitigate this risk. For instance, Facebook maintained a secret program known as XCheck to exempt high-profile accounts from their routine content moderation programs, with the explicit goal of avoiding errors that might result in scandal (Horwitz, 2021).

In this light, platforms may also be inclined to *exaggerate* the technical importance of shadow banning in their moderation strategies, as a means to combat 'bad actors'. Their claims here should be taken with a grain of salt. The trade-offs between due process and efficacy in content moderation

may not be so costly as platforms themselves suggest. For platforms, shadow banning doesn't just outwit bad actors; it avoids bad press.

Overall, then, the incentives toward shadow banning are several. Its most important driver may be the general turn to visibility remedies, as a response to disinformation and other recent controversies around "lawful but awful" content. These new techniques result in less observable moderation decisions, and therefore lend newfound significance to official notices as transparency safeguards, not just to explain moderation decisions but to notify them. In this way, the turn to visibility remedies offers a pretext for platforms' more general tendency toward secrecy in content moderation, which is driven by both technical and political considerations. All these factors suggest that shadow banning will likely persist unless platforms face sufficient pressure to end it. Enter: the Digital Services Act.

3. Transparency rules for content moderation in the Digital Services Act

The Digital Services Act (DSA) is not the first legislation to regulate transparency in content moderation, but it is the first to address the issue of shadow banning directly.[2] The following section proceeds by introducing the general features of the DSA's notice-and-action framework for content moderation, including its definition of shadow banning in Recital 55. It then highlights two key provisions that regulate shadow banning practice: Article 14 on Terms of Service, and Article 17 on the Statement of Reasons.

3.1. *The DSA's notice-and-action framework for content moderation*

The DSA is a lengthy and complex piece of legislation, but it is fair to say that its main concern is content moderation. This it regulates in three main ways. First, it restates, with only minor revisions, the pre-existing 'safe harbour' regime governing liability for unlawful user-generated content.[3] Second, it outlines a comprehensive procedural framework for all content moderation actions, known as the 'Notice-and-Action' framework. What makes this framework especially novel is that it applies not only to content prohibited by law, but also to content prohibited under the service's own terms and conditions. Third, the DSA sets out duties of care for the very largest platforms, known as 'systemic risk mitigation'. Most relevant for our purposes is the second element: notice-and-action.

The DSA's notice-and-action framework applies to all content moderation actions, a concept which it defines broadly. Earlier content moderation laws have concerned themselves almost exclusively with content removal and account suspension, in what Eric Goldman has termed the 'binary leave up / take down paradigm' (Goldman, 2021). The DSA innovates with a more expansive understanding of content moderation actions that expressly includes non-removal remedies such as demonetisation and visibility restrictions (Article 3(t) DSA). Recital 55 clarifies what the DSA means by visibility restrictions, and even mentions shadow banning explicitly:

> *Restriction of visibility may consist in demotion in ranking or in recommender systems, as well as in limiting accessibility by one or more recipients of the service or blocking the user from an online community without the user knowing it ('shadow banning').*

This recital clearly uses shadow banning in the original, narrow sense of secret account suspensions, rather than the modern, broad sense of secret visibility reductions. From a legal standpoint this matters little, however, since the phrase "shadow banning" is only used in this recital and does not return in the DSA's actual enacting provisions (i.e. its "articles"). Going forward, lawyers would do well to keep in mind this gap between statutory and popular usage. But regardless of these semantics, the fact remains that visibility remedies, which attract the bulk of shadow banning speculation, are recognised as content moderation actions, and are therefore subject to the DSA's notice-and-action procedures.

The DSA's notice-and-action framework operates as follows. Its cornerstone is Article 14 DSA, which lays down two key principles: First, the rules governing online intermediaries' content moderation must be published in their Terms and Conditions, in "clear and unambiguous language". Second, these rules must be enforced "in a diligent, objective and proportionate manner", and with due regard to the interests and fundamental rights involved (q.v. Appelman et al., 2021). Article 16 adds that these services must offer a notice mechanism through which third parties can flag content for content moderation review. Crucially for our purposes, Article 17 requires that online intermediaries provide a Statement of Reasons to the affected uploader for each content moderation decision (regardless of whether the action is taken in response to a notice or on the service's own initiative). These actions must also be open to appeals through internal complaint handling (Article 20) and through external dispute resolution (Article 21). Taken together, this framework reflects the basic principles of due process: every sanction – i.e. any deprivation of lawful interests – must be governed by clear and foreseeable rules; must be notified and explained to the affected users; and must be open to reasoned appeals (Suzor, 2018). As we will see below, this leaves little room for shadow banning.

This is only a basic sketch. The DSA introduces many more transparency rules besides, but these are generally less relevant to the issue of shadow banning. For instance, the DSA also contains public reporting requirements for content mod-

[2] The platform-to-business regulation or 'P2B Regulation' contains a similar set of rights in Articles 3 and 4. This instruments was adopted in 2019, only three years before the DSA. This contribution focuses on the DSA since its rights are both deeper in substance and broader in scope; the DSA's relevant safeguards apply to all users of hosting services, including platforms, whereas the P2B Regulation applies only to business users of online intermediation services (See P2B Regulation, Articles 3 and 4). The DSA is also the first EU legislation to expressly use the term 'shadow ban'. Most shadow banning cases covered by the P2B Regulation are therefore covered by the DSA's rules as well, whereas the inverse is not true. Some additional comparative reflections are included at the end of this section.

[3] Its predecessor is Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('e-Commerce Directive').

eration actions (e.g. Articles 15, 23, and 42), explanation duties for recommender systems (Article 27), and data access for regulators and researchers (Article 40). Yet these provisions are not designed to regulate individual moderation actions or shadow bans.

3.2. *Article 14 DSA on terms and conditions*

Article 14(1) DSA demands that platforms codify their content moderation rules. In "clear and unambiguous language", their Terms and Conditions must set out information about the restrictions they impose regarding user-generated content. This disclosure "shall include information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review." I refer to this as the codification principle, since it reflects the rule of law principles of legality, foreseeability, and accessibility for the imposition of sanctions.

This codification principle is relatively novel in platform regulation. In keeping with the binary paradigm's focus on unlawful content (Goldman, 2021), most content moderation laws have heretofore left platforms' internal rules largely unregulated.[4] Precursors to the DSA's codification principle can already be found in national court precedents, which have placed limits on overly vague Terms based on fundamental rights, consumer protection, and general principles of private law (e.g. Ketteman and Tiedeke, 2020). In this light, Article 14 DSA's codification principle is not entirely new, but instead serves to clarify, and perhaps strengthen, the pre-existing duty for platforms to stipulate clear and specific content moderation policies.

Most major platforms already publish content policies, and these have become more detailed over time. Still, these voluntary efforts continue to be criticised for their lack of detail, and Article 14 DSA might force further reforms by holding them to its standard of clear and unambiguous language. Its impact may be especially significant for non-takedown remedies, which tend to be given short thrift in platforms' current Terms, being governed by relatively generic policies such as restrictions on "inappropriate" or "borderline" content (Heldt, 2020; Horten, 2021). Facebook recently published a systematic overview of its (down)ranking policies, known as its Content Distribution Guidelines, but this is the exception to the general rule that the policies for most non-takedown remedies are not published in the same systematic detail as for takedown (Facebook, 2020). Article 14 DSA would demand a more systematic and comprehensive approach to such documentation for all online platforms and for all moderation measures, and help to shed light on visibility management policies.

Still, Article 14 DSA is only a partial solution for shadow banning. If enforced properly, it might provide some reassurances by improving the foreseeability of platform policies and helping users to self-assess their compliance. For this to succeed, the Terms would need to be both detailed and clear, and even then, it would only help relatively sophisticated and proactive users—those with the wherewithal to seek out and study these disclosures. Most users, we know, do not consult Terms and Conditions. But experts do (Mahieu and Ausloos, 2020; Fung, 2013).

Even towards experts, there is little cause for optimism about the foreseeability that Terms can provide. Like all contracts, platform Terms face the basic problems of indeterminacy and contractual incompleteness; no statute or contract is ever sufficiently detailed to cover all contingencies, and will inevitably leave room for interpretation. Even legal doctrines with centuries of jurisprudence behind them, such as defamation or fair use, continue to divide lawyers, leaving little hope that enforcement of platform Terms should ever be any more foreseeable. Indeed, excessively detailed codifications may not even be desirable due to tradeoffs with flexibility and substantive fairness, which could unduly hamper moderators in unforeseen circumstances.

Adding to this challenge of foreseeability are the practical constraints of content moderation at scale. Content moderation is not a process of careful legal-professional reasoning, but an industrial process that occurs at massive scales through standardised routines (Roberts, 2019). In light of its massive scale, Evelyn Douek (2022) proposes that content moderation is best understood as an administrative bureaucracy rather than as a judiciary carefully weighing individual cases. And even this administrative analogy, as Douek herself acknowledges, may overstate the role of human judgement. Human moderators, if at all involved, are typically forced to decide on moderation actions through snap judgements and crude heuristics, and rarely have time for careful deliberation or fact-finding (Roberts, 2019). For instance, Facebook instructed its moderators to classify content as terrorist propaganda for the mere *mentioning* of certain terrorist organisations (Biddle, 2021). Many more decisions are automated entirely (Roberts, 2019; Bloch-Wehba, 2020). Relying on automated machine-learning classifiers, these automated decisions operate through statistical inferences bearing little or no resemblance to human reasoning as expressed in language-based rules.[5] The true drivers of content moderation, therefore, are often far removed from the policy principles that nominally govern them.

For all these reasons, Article 14's Terms and Conditions contribution to foreseeability is likely to be modest. Its most

[4] Some exceptions: Variants on the DSA's codification principle – though far more narrowly tailored - can be found in recent sectoral frameworks such as the recent Platform-to-Business Regulation (as regards the ranking of business users by *ecommerce platforms*) and the Audiovisual Media Services Directive (as regards the protection of minors, hate speech and terrorism on video sharing platforms). See: Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services, Article 3. Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities 2018, Article 28(b)(3)(a).

[5] Amélie Heldt (2020) cites the telling example of Facebook misclassifying a pair of onions as due to the 'overtly sexual manner' they were positioned—evidently the result of a machine-learning classification error.

important function may be not as an *ex ante* guide to user conduct but rather as an *ex post* rubric for appeals and error correction; a form of *justificatory* transparency which aims to establish and help vindicate individual rights (Kaminski, 2020). Of course, the problem with shadow banning is that it precludes all opportunities for such appeals and error corrections. For shadow banning, Article 14 may provide some minimal opportunity for self-assessment, but since errors and ambiguities cannot be ruled out it does not address the issue directly. To even begin evaluating platforms decisions, users must first be made aware of them. For that, we must turn to Article 17.

3.3. *Article 17 DSA on the statement of reasons*

Article 17 DSA demands that each moderation action be accompanied by a "Statement of Reasons" to the affected user. This statement must include the following information: (1) the measure taken; (2) the legal or contractual violation that this measure responds to; (2) the facts and circumstances relied on in taking the decision; (3) information on the role of automated decision-making in this action, (4) whether or not the measure was taken in response to a third party notice; and (5) the user's possibilities for redress.

Article 17 fulfils at least two distinct functions: notification and explanation. Notification makes users aware of sanctions, whereas explanation aims to give reasons for those sanctions. Explanation is a crucial feature of due process, and raises many difficult policy questions in the context of (automated) content moderation (Gorwa et al., 2020). But these are tangential to the issue of shadow banning, which, as discussed, is primarily a problem of notification. Indeed, if shadow banning is characterised by a lack of notification, then Article 17's notification duty can be read as a *prohibition* on shadow banning.

Article 17 does contain exceptions, however. First, it does not apply to moderation actions taken in response to removal orders by public authorities, as regulated under Article 8 DSA. This exception is not immediately relevant to the problem of shadow banning, since it only applies to removals and not visibility remedies. More importantly for our purposes, Article 17 DSA also exempts content moderation actions affecting "deceptive high-volume commercial content". This exception is worth discussing in detail, as it is here that the DSA attempts to balance the competing interests at stake in shadow banning.[6]

This clause about high-volume commercial content seems to envision a narrow exception for shadow banning in the context of advertising spam. That the EU legislator should side with secrecy here, stands to reason; advertising spam is perpetrated by relatively persistent and well-resourced adversaries, and appeals to no significant public interests. In advertising spam, therefore, the public interest in transparency and due process is relatively low, and the public interest in secrecy (and thus the effective combating of adversarial spammers) is relatively high. A broader exemption might also have included political spam, in what is known as "information operations" or "coordinated inauthentic behaviour" (François and Douek 2021; Giglietto et al. 2020). But the DSA's focus on commercial content suggests that such political activity is too sensitive from a public interest perspective to permit unaccountable shadow banning.

More surprising is the proviso that this commercial content must be "deceptive" for shadow banning to be permitted. This is a substantial narrowing. After all, spam can be unwelcome even when it is factually accurate. And for platforms to check for truthfulness in user content is a major operational burden, since these services must moderate many millions or even billions of such items every year. It is also unclear how this exception will apply to moderation actions taken against *accounts* or *users*, given that the exception refers to deceptive commercial *content*. Overall, then, the exception is relatively narrow, and leaves little room for shadow banning at all.

The DSA's secrecy rules can be contrasted with those in the P2B Regulation, which offer comparable content moderation transparency rights for business users. Here the exceptions are generally broader and more flexible. First, the P2B Regulation's statement of reasons in Article 4 takes an actor-based approach, and simply permits secrecy in cases where the business user in question "has repeatedly infringed the applicable terms and conditions" (P2B Regulation, Article 4(5)). This actor-based approach will likely appeal to platforms since it is far more practicable to assess repeat violations than to assess veracity. Then again, if interpreted too broadly, the concept of "repeat infringement" does risk restricting due process for ordinary users acting in good faith.[7] Second, the P2B Regulation's disclosure rules for ranking in Article 5 attempt to manage security and circumvention concerns by introducing an exemption for the disclosure of "any information that, with reasonable certainty, would result in the enabling of deception of consumers or consumer harm through the manipulation of search results" (P2B Regulation, Article 5(6)). This exception based on the *substance of disclosures* seems to enable platforms to modulate the level of detail given in explanations, without necessarily impinging on the basic, prior safeguard of notification. In sum, whereas the DSA's secrecy rules focus on the nature of the moderated *content*, the P2B shows how considering the *actors* and *disclosures* might also be relevant parameters for the balancing of transparency against secrecy. In this light, the DSA's shadow banning exceptions are not only narrow but somewhat inflexible, in that they focus only on the moderated content and do not take into account other factors.

Another factor that might be considered is the nature of the *enforced rule*. For instance, actions against child sexual abuse imagery or cyberstalking might justify a greater degree of secrecy than those against clickbait or conspiracy theories. As to account-based factors, besides repeat infringers, one might also consider an exemption for new accounts; rapidly creating

[6] The DSA's legislative history supports this reading; in the original proposal, the Statement of Reasons applied only to takedown decisions, and did not contain any exemptions for commercial content. The same round of amendments which expanded this provision to cover all moderation actions, also added the exemption for high-volume commercial content.

[7] For instance, YouTube's 'copyright strikes' systems escalates sanctions users as little as three violations over as long as a six-month period. This approach has been criticised for the risk of chilling effects on user activity. See: Bridy (2020), citing Wodinsky (2019).

new accounts is an important strategy for spammers to circumvent account suspensions and terminations. But for the DSA, a brand new account with zero followers or post history seems to be entitled to the same due process treatment as an established pillar of the community. Clearly, the cost-benefit analysis for due process is complex and may vary significantly across all these different cases.. But for Article 17 DSA, all that matters is whether the item contains high-volume deceptive commercial content.

More fundamentally, the DSA's approach is inflexible in that it bundles all relevant due process rights—notice, explanation and appeals—into the singular concept of a 'moderation action'. In practice there may be a large set of edge-cases where integral explanation and/or appeal could be onerous in terms of costs, or too sensitive in terms of security, but where a bare notice right could still be of substantial value as a bulwark against shadow banning and as a minimal precondition for legal and social accountability. In this light, the DSA's attempt at balancing is somewhat rudimentary, and in future may benefit from further refinement, such as by incorporating more factors into the shadow banning calculus and unbundling notice safeguards from other aspects of due process.

At present, the DSA's rigid design still deserves praise for erring on the side of transparency rather than secrecy, and thereby providing an impetus for more informed debate on the merits of shadow banning. Until now, the case for shadow banning has rested primarily on untested technocratic arguments about circumvention. As I have argued in Section 2.3 above, these claims are not only difficult for outsiders to assess but also risk giving cover to the platforms' more general disinterest in accountability and due process. The DSA, by erring on the side of transparency, will put these arguments to the test, forcing platforms to demonstrate the practical need for shadow banning (if any) and make these claims available for public scrutiny. If greater secrecy is deemed necessary in future legislation, then this will at least be a secrecy arrived at through public rulemaking, rather than, as present, a secrecy taken on faith from self-interested platforms.

Regardless of such future revisions, a more pressing practical problem for the DSA's enforcement is how it defines moderation actions in the first place; what it means for an item to moderated. As I will discuss below, the category of ranking or 'demotion' sanctions is especially problematic.

4. Ranking due process between moderation and curation

4.1. *Defining demotion and the problem of counterfactuals*

Compared to most content moderation remedies, it is not so clear what it means to 'demote' an item. Most other remedies can be summed up in relatively straightforward binaries: an item can either be left up or taken down; listed or delisted; monetised or demonetised; an account active or suspended. But when is an item 'demoted' or 'downranked', as opposed to merely 'ranked'? The basic problem is that ranking is a zero-sum process in which all items receive differential treatment, leaving no clear baseline of ordinary or default treatment for comparison. In other words, demotion lacks a clear counterfactual.

Several commentaries have already remarked on this problem of counterfactuals as an obstacle in regulating ranking moderation. For Rachel Griffin (2022), it counsels against a human rights approach to ranking governance: ranking interventions are "difficult to frame as a clear-cut rights violation", since, after all, "[*w*]hat level of algorithmic visibility does anyone have a right to?" Gillespie (2022) concludes that "it is nearly impossible to be transparent about reduction policies", since, after all, "[*h*]ow does one measure or document reduction: what should the reduced visibility of a piece of content be compared to?" Very similar objections have also been raised against the regulation of "amplification", which refers to excessive visibility rather than restricted visibility and in this sense can be seen as the mirror image to demotion. Daphne Keller (2021) objects that proposals to regulate amplification are "hard to assess, because it is hard to define", and for Luke Thorburne, Jonathan Stray and Priyanjana Bengagina (2022) the concept of amplification is "not precise enough to be used in law". This problem of counterfactuals poses a definitional challenge for the DSA's regulation of demotion. It also speaks to an ambiguity in the shadow banning imaginary itself: both imply some underlying distinction between ordinary ranking routines and exceptional ranking sanctions.

I want to offer a slightly more optimistic account. Without denying the problem of counterfactuals, I propose that a workable legal concept of "demotion" might still be devised through detailed engagement with specific ranking architectures. Demotion practices come into view more clearly when one recognises that the platform ranking process does not consist of one single, monolithic Algorithm, but is instead comprised of many fragmentary organisational and computational units all working in concert but fulfilling distinct functions (Rieder and Hofmann, 2020; Seaver, 2017). In these complex assemblages, it is possible to distinguish certain subsystems that ascribe relevance scores to content (typically optimised for user engagement), and others that impose *ex post* maluses or bonuses on these scores based on ulterior optimisation processes, such as clickbait or hate speech classifiers. In other words, certain subsystems *produce* algorithmic relevance scores, whereas other subsystems serve only to *reduce* them (Gillespie, 2022). The former optimises for engagement, the latter for compliance. It is these reduction decisions that most clearly constitute moderation actions. This interpretation manages the problem of counterfactuals by taking engagement optimisation as its baseline treatment, against which reductions can then be defined.

Facebook's own description of its Newsfeed ranking process (Fig. 1 below) can serve as an example. It includes three main steps involving different sets of machine learning classifiers: (1) inventory or candidate generation, which selects several hundreds of possibly relevant candidates out of the pool of available content, (2) relevance scoring, which attributes initial ranking scores to all candidates based on 'multitask model' for engagement optimisation, and (3) integrity processes, which test items for compliance with rules such as those on borderline content and spam. Whereas steps (1) and (2) appear to optimise for relevance, and together pro-

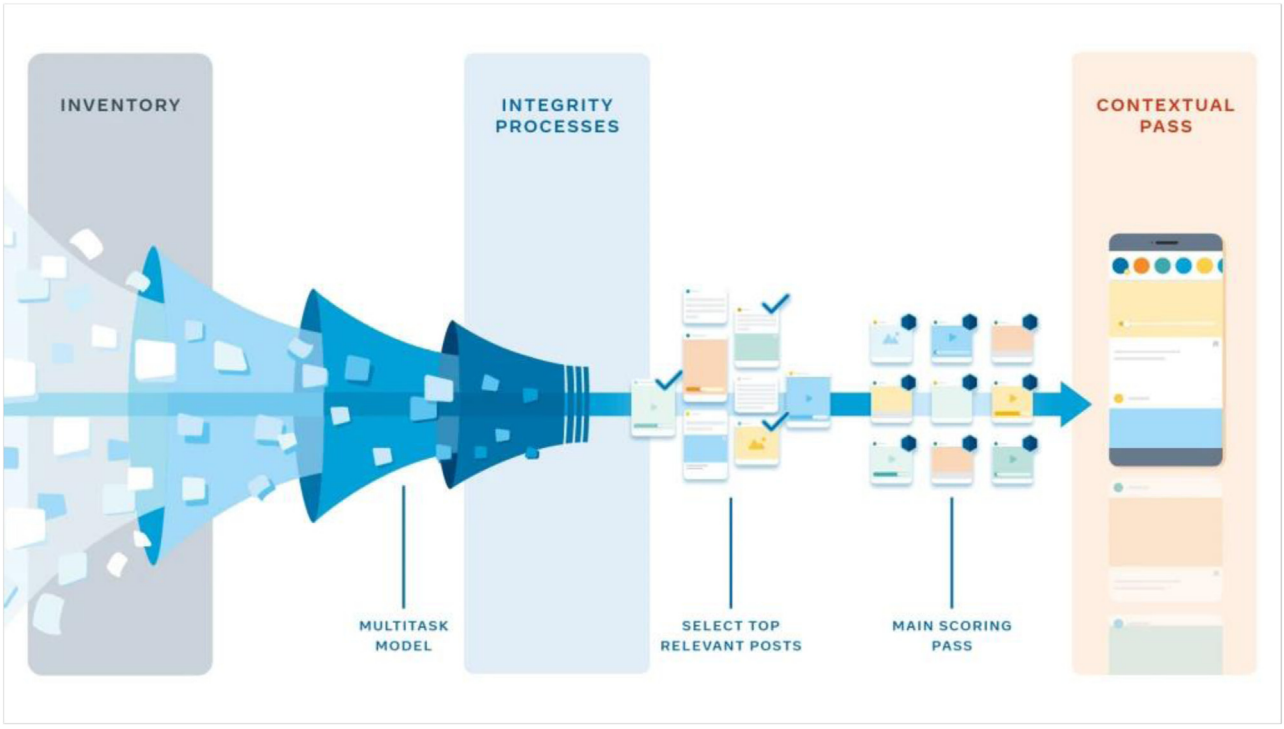

Fig. 1 – Facebook's schematisation of its Newsfeed ranking procedure.[8]

duce relevance scores, step (3) optimises for entirely different classifiers, often content-related, to reduce relevance scores. These integrity processes, then, might result in 'demotions' for purposes of EU law.

Returning to the problem of counterfactuals, all this means that platforms should be able to disclose *whether* an item has been demoted. What remains challenging, however, is to determine *by how much* an item has been demoted. Since content ranking is user-driven and path-dependent, it is all but impossible to determine how an item would have performed had it not been demoted (Gillespie, 2022). Estimates might be made by observing the *average* impact of a given demotion technique, for instance by comparing demoted and non-demoted items from individual accounts, but these would only be rough estimates. From a legal perspective, these questions around the impact of demotion could be relevant in assessing proportionality or damages related to wrongful demotions. But this calculation problem is less germane to the issue of shadow banning, which, as discussed, revolves around the prior question of *whether* an item has been demoted, and for which the question of impact remains secondary.[9]

EU law, therefore, appears to be coalescing around a concept of demotion as an *ex post* reduction of engagement scores. This approach is open to critique, however. Models such as Facebook's focused on "integrity processes" risk concealing other interventions in the system and other exercises of ranking power, insofar as it does not account for the ways in which platforms govern visibility through the relevance scoring process itself. "When we are fighting about particular dynamics of virality", Tarleton Gillespie (2022) warns, "we are not asking whether there are other logics of circulation that we should prefer". Further to this point, it is worth noting that relevance scoring is not a fixed or objective process but one that is itself iterative and political. Constructs such as "engagement", "relevance", or "quality" may seem objective, but in practice their measurement entails a complex and value-laden weighing of competing interests (Van Couvering, 2007; Van Hoboken, 2012; Gillespie, 2014; Napoli, 2015; Helberger, 2019; Leerssen, 2020). Platforms act as gatekeepers not just by ruling on exceptions to the ranking game, but by writing the rules to this game and revising these over time (Cotter, 2021). Relevance scoring may therefore harbour its own forms of content regulation, which 'demotion' safeguards would then fail to capture.

An example to illustrate this point is the history of Facebook's reaction feature, as reported by the Washington Post (Merill and Oremus, 2021). The 'Like'-button has long been an important component of Facebook's engagement optimisation metrics, but in 2016 the platform added several new

[8] This is a screenshot taken from Facebook's official website. The 'contextual pass' mentioned in this schema refers to an additional step accounting for contextual considerations such as content diversity. See: Lada et al. (2021).

[9] Admittedly, these questions of *whether* and *how much* may still overlap insofar as one might try to impose a *de minimis* standard for demotion due process. For instance, one might argue that light-touch interventions resulting in a mere 1% reduction are insufficient to raise due process concerns as moderation sanctions, and should therefore be exempted from the DSA's procedural safeguards. But since calculating the precise impacts of any measure is so complex, the problem of counterfactuals counsels against any such quantitative thresholds.

options including a 'Haha', 'Wow', and 'Angry' react. In order to encourage users to experiment with these new and unfamiliar features, Facebook initially measured these new reacts as a stronger form of engagement than a conventional 'Like'. Later, the platform observed that the 'Angry' emoji correlated strongly with low-quality content and disinformation. To slow the spread of this content, the platform reduced the engagement signal of Angry reacts to zero. In this way, Facebook suppressed content not by reducing its relevance scores, but instead by changing how they define relevance in the first place.

In this sense, a regulatory project focused on "demotion" risks overlooking the structural (Griffin, 2022) or constitutive (Cotter, 2021) aspects of platform ranking power. Still, it has the advantage of singling out relatively fine-grained and targeted interventions. Structural changes to engagement optimisation, such as Facebook's reaction feature changes, struggle to single out specific targets, and can only do so indirectly based on observed patterns in user engagement. But ex post reductions, by contrast, afford a relatively fine-grained form of control. They permit platforms to curate content not only by tweaking relevance metrics but on wholly separate criteria, including automated but also manual human intervention. From a freedom of expression perspective, therefore, these more targeted demotion sanctions may arguably raise heightened concerns of censorship or viewpoint discrimination; they provide a venue for platforms to exercise content-specific "opinion power" (Helberger, 2022) or "curatorial power" (Poell et al., 2022) in ways that the engagement optimisation process itself may not. In this sense, ex post restrictions raise distinct risks from a fundamental rights and due process perspective, which arguably require distinct safeguards.

In light of the above, I conclude that it is not entirely futile or incoherent to regulate demotion as a category of content moderation sanctions under the DSA. It is, however, technically complex and normatively incomplete as a means of regulating ranking power. Especially in light of regulatory agencies' limited technical capacities, this complexity may provide platforms with occasions for obfuscation. Transparency in practice is performative, and alters the practices it documents (Flyverbom, 2019). In the same way that public meeting rules push lawmakers into backchannels, Article 17 DSA might encourage platforms to hide their most controversial measures beyond those sites which the law has recognised as "content moderation". For these reasons, platforms' descriptions of their ranking process cannot be taken at face value. In order to determine the mechanisms of 'demotion, regulators will need to take full and independent stock of platform ranking procedures.

Even if Article 17 DSA is enforced rigorously, and all demotion is disclosed dutifully, what it probably cannot do is put an end to shadow banning *suspicions* and *allegations*. Users will continue to face sudden and inexplicable drops in visibility, if not due to secret *ex post* sanctions then due to more systemic *ex ante* adjustments to the ranking system; or simply due to the ever-shifting whims of audience taste and attention. This precarity is a structural feature of ranking (Duffy, 2020). From the user perspective, these fluctuations may be functionally indistinguishable from shadow banning, and will likely continue to arouse suspicions of foul play. That the law does not recognise users' rise and fall as a result of 'content moderation', may be of little reassurance to them. Helping publics to grapple with these more constitutive dimensions of ranking power demands that we move past narrow concerns with shadow banning and content moderation sanctions, and towards a more comprehensive reckoning with the precarities of content curation.

4.2. *Ranking transparency beyond the downrank: from moderation to curation*

The above has shown that important aspects of ranking governance cannot be broken down into individual cases of content moderation—into discrete demotion sanctions depriving specific individuals of their lawful interests. The structural or constitutive features of content ranking are integrated into the engagement optimisation process; they *produce* ranking rather than merely reducing it. Addressing these demands a more expansive approach to transparency and accountability in ranking systems, not only as an occasional site of content moderation but as structural site of content curation. What new models of ranking transparency come into view when we look beyond the restrictive categories of content moderation, demotion, and shadow banning?

A first step, already taken in the DSA, is Article 27's codification principle for recommender systems, which is separate from its codification principle for moderation policies. Article 27 DSA requires platforms to disclose "the main parameters used in their recommender systems", including "the criteria which are most significant in determining the information suggested to the recipient of the service." In addition, this provision requires that platforms state "the reasons for the relative importance of those parameters." Since this provision is not limited to moderation sanctions, it can start to shed light on curation in a more comprehensive sense without being confined to "demotion".

General codification rules such as these still face important limitations, however. Abstract descriptions of recommender algorithms struggle to shed meaningful light on their operation in practice, due to the extreme complexity of their machine-learning algorithms as well as the sociotechnical contingency of their interaction with user content and audiences (Leerssen, 2020). Furthermore, Article 27 does not even require in-depth explanations; all it asks for is a description of "main parameters". At worst, these descriptions could be so generic as to offer no practical guidance. But assuming a more robust implementation, it might still function as a useful complement to individual content moderation transparency: when users experience a sudden drop in traffic, and receive no notice of individual moderation actions under Article 17 DSA, this might then prompt them to check for general updates to curation policies under Article 27 DSA. It goes beyond the scope of this paper to explore these possibilities in detail, but a robust ranking disclosure might for instance elaborate on the following: the different engagement signals and other data sources used; the relative weighting of different engagement metrics and other relevance signals; possibilities for users to alter their recommendation experiences; and information about the types of content and formats which tend to perform well – all in addition, of course, to the various visi-

bility reduction policies in place (Bengani et al., 2022).[10] Ideally, these disclosures would include a changelog so that important changes and updates could be tracked over time. The appropriate level of detail here may depend on the size and sophistication of the platform in question.[11]

More ambitious reforms would focus on access to ranking data. Concerns about shadow banning are fuelled by a lack of granular traffic data, which prevents uploaders from observing their performance in ranking systems in detail. The available data is often limited to view and engagement aggregates, with little information offered on actual recommendation trends and audience discovery pathways – or reserved only for paying customers. Expanding access to this data could serve a dual function. First, access to analytic data could help to enforce the DSA's prohibition on shadow banning, by helping to detect undisclosed instances of demotion. Without this data, shadow banning will remain a known unknown, and its enforcement could be even more likely to fail. Second, access to analytic data could help users and publics to understand curation trends in a broader sense going beyond mere demotion or moderation, shedding light on the impacts of structural policy choices and sociotechnical dynamics of ranking curation. Ideally, such data would also be made available not only to uploaders themselves but also to other stakeholders in platform governance, and to the public at large (Leerssen, 2020; Rieder and Hofmann, 2020). In designing such access frameworks, the public interest in transparency would have to be weighed against competing interests in privacy, service security, and (to the extent their business model is considered worth protecting) the commercial interests of platforms.

5. Conclusion

Visibility remedies are making content moderation more nuanced, but less transparent. The blunt instruments of content takedown and account suspension were always largely self-evident in their effects. But visibility remedies leave barely any trace, since they play out through dynamic and volatile ranking systems which serve to obfuscate their effects. Recent allegations of shadow banning can be understood as a response to these new moderation techniques, and how these lend newfound significance to the question of transparency rights in content moderation.

The DSA, as the first major legislation to regulate visibility remedies, makes shadow banning a legal problem. Its due process notice requirements, I have shown, amount to a general prohibition on shadow banning, with only narrow exceptions for high-volume deceptive commercial content. This approach leaves relatively little flexibility to balance the competing interests at stake in content moderation secrecy, particularly as regards non-deceptive and non-commercial forms of high-volume spam. I have argued that future revisions may require a more nuanced set of exceptions, based not only on the affected content but also taking into account other factors such as the actors and norms at issue. Unbundling the due process rights of notice, explanation and appeal may also help in striking this balance. Although the DSA may lack nuance on these points, its choice to err on the side of transparency appears sensible, since it helps to bring these balancing considerations out in the open. The case for transparency is already clear, but the case for shadow banning remains speculative and undependable—difficult for outsiders to assess and tempting for platforms to exaggerate. By erring on the side of transparency, the DSA places the onus on platforms to demonstrate the practical importance of shadow banning (if any!) and make it available for public scrutiny. Should future lawmaking opt for a return to shadow banning, then this will at least be a secrecy arrived at through public rulemaking, rather than the present secrecy taken on faith from self-interested platforms.

The final section of this paper has highlighted a more fundamental problem: the meaning of "demotion" as a category of moderation sanctions. This concept is central to the shadow banning imaginary and the DSA's response to it, and yet its meaning is far from straightforward. Building on earlier criticism around the problem of counterfactuals in ranking regulation, I have argued that demotion is not necessarily incoherent as a legal category of moderation sanctions, if understood as an *ex post* modification to content relevance scores. Understood in this way, safeguards against demotion may help to shed light on relatively fine-grained and targeted exercises of platform opinion power. Still, it should be kept in mind that these demotion safeguards do not account for the constitutive aspects of ranking governance, not only as moderation but also as curation; how platforms govern visibility not just by ruling on ranking exceptions, but by writing and constantly revising the rules of the ranking game.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

[10] There is also an extensive literature on algorithmic transparency in other domains. Its relevance here is limited here to insofar as it tends to focus on explaining highly consequential individual decisions, such as criminal profiling and credit scoring. Recommender systems, by contrast, operate at a rapid pace, offering dozens, hundreds or thousands of recommendations to most users every day, leading to a comparatively limited interest in exhaustively explaining individual decisions and a correspondingly greater emphasis on systemic questions of composition and diversity across aggregates. With this caveat, further guidance might be sought for instance in GDPR-related discussions on the 'right to an explanation' (e.g. Edwards and Veale, 2017; Selbst and Powles 2017; Kaminski and Malgieri 2020).

[11] Article 27, like Article 14, applies to all platforms regardless of size (exempting only small- and micro-sized enterprises), and an overly demanding interpretation could be onerous on smaller players. A higher standard for the largest platforms could arguably be derived from the systemic risk framework in Articles 33 and 34, which requires the largest platforms to assess and mitigate systemic risks to users' right to expression and information, in particular as regards "the design of their recommender systems".

Acknowledgements

Redacted for peer review.

REFERENCES

Appelman N, Quintais J, Fahy R. Using terms and conditions to apply fundamental rights to content moderation: is article 12 DSA a paper tiger?. Verfassungsblog; 2021. 10 September 2021. Available at: https://dsa-observatory.eu/2021/09/10/using-terms-and-conditions-to-apply-fundamental-rights-to-content-moderation-is-article-12-dsa-a-paper-tiger/(accessed 28 June 2022) [https://perma.cc/DW7Y-BQYQ].

Are C. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. Feminist Media Studies 2021:1–18.

Bengani P, Stray J, Thorburn L. A Menu of Recommender Transparency Options. Tech Policy Press; 2022. Available at: https://techpolicy.press/a-menu-of-recommender-transparency-options/(accessed 4 January 2023) [https://perma.cc/N8D6-452A].

Biddle S. Revealed: facebook's secret blacklist of "dangerous individuals and organizations. The Intercept; 2021. Available at: https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/(accessed 28 June 2022) [https://perma.cc/ZM92-45UB].

Bloch-Wehba H. Automation in moderation. Cornell Int Law J 2020;53:41.

Bodo B, Helberger N, Irion K, Zuiderveen Borgesius F, Moller J, van de Velde B, et al. Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents. Yale J Law Technol 2017;19(1):133–81.

Bridy A. The price of closing the value gap: how the music industry hacked EU copyright reform. Vanderbilt J Entertain Technol Law 2020;22:323.

Cobbe J. Algorithmic censorship by social platforms: power and resistance. Philos Technol 2021;34(4):739–66.

Cotter K. Playing the visibility game: how digital influencers and algorithms negotiate influence on Instagram. New Media Soc 2019;21(4):895–913.

Caplan R, Gillespie T. Tiered governance and demonetization: the shifting terms of labor and compensation in the platform economy. Soc Media+ Soc 2020;6(2).

Cotter K. Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. Inf, Commun Soc 2021:1–18.

Directive (EU) 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in The Internal Market ('e-Commerce Directive').

Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities 2018.

Diresta R. Free speech is not the same as free reach. Wired Magazine; 2018. Available at: https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/(accessed 28 June 2022)[https://perma.cc/7DR5-3S8W].

Douek E. Content moderation as administration. Harv Law Rev 2022;136 forthcoming.

Edwards L, Veale M. Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. Duke Law Review 2017;16:18–84.

Facebook (2020) 'Sharing our content distribution guidelines'. Available at: https://about.fb.com/news/2021/09/content-distribution-guidelines/(accessed 28 June 2022). [https://perma.cc/BRT3-7XC8].

Flyverbom M. The Digital Prism: Transparency and Managed Visibility in a Datafied World. Cambridge University Press; 2019.

François C, Douek E. The accidental origins, underappreciated limits, and enduring promises of platform transparency reporting about information operations. J Online Trust Safety 2021;1(1).

Fung A. Infotopia: unleashing the democratic power of transparency. Polit Soc 2013;41(2):183–212.

Gorwa R, Binns R, Katzenbach C. Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc 2020;7(1).

Goldman E. Content moderation remedies. Michigan Technol Law Rev 2021;28:1.

Giglietto F, Righetti N, Rossi L, Marino G. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. Inf, Commun Soc 2020;23(6):867–91.

Gillespie T. Do not recommend? Reduction as a form of content moderation. Soc Media+ Soc 2022;8(3).

Gillespie T. The relevance of algorithms. In: Gillespie Tarleton, Boczkowski Pablo J, Foot Kirsten A, editors. Media Technologies: Essays on Communication, Materiality, and Society. MIT Press; 2014. p. 167.

Griffin, R. (2022) 'Rethinking rights in social media governance: human rights, ideology and inequality', SSRN Draft Paper. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064738 (accessed 4 January 2023).

Helberger N. On the democratic role of news recommenders. Digital Journalism 2019;7(8):993–1012.

Helberger N. The political power of platforms: how current attempts to regulate misinformation amplify opinion power. Digital J 2022;8(6).

Heldt A. Borderline speech: caught in a free speech limbo?. Internet Policy Review; 2020 Available at: https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510 (accessed 15 July 2022).

Horten M. Algorithms patrolling content: where's the harm? An empirical examination of Facebook shadow bans and their impact on users. SSRN Draft Paper; 2021 Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792097 (accessed 28 June 2022).

Horwitz J. Facebook says its rules apply to all. Company documents reveal a secret elite that's exempt. Wall Street J 2021. September 13. Available at: https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353. (accessed 28 June 2022).

Jaidka K, Mukerjee S, Lelkes U. Censorship on social media: the gatekeeping functions of shadowbans in the American Twitterverse. SSRN Draft Paper; 2022 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4087843 (accessed 28 June 2022).

Kaminski M. Understanding transparency in algorithmic accountability. In: Barfield Woodrow, editor. Cambridge Handbook of the Law of Algorithms. Cambridge University Press; 2020.

Kaminski M, Malgieri G. Multi-layered explanation from algorithmic impact assessments in the GDPR. Proceedings of FAT* '20, January 27–30, 2020; 2020. p. 68–79.

Keller D. Amplification and its discontents: why regulating the reach of online content is hard. J Free Speech Law 2021;1:227–68.

Kettemann MC, Tiedeke AS. Back up: can users sue platforms to reinstate deleted content? Int Policy Rev 2020;9(2).

Lada A, Wang M, Yan T. How does news feed predict what you want to see?. Meta Newsroom; 2021 Available at: https://perma.cc/N2NP-C6CD (accessed 28 June 2022).
Le Merrer E, Morgan B, Trédan G. Setting the record straighter on shadow banning. IEEE INFOCOM2021-IEEE Conference on Computer Communications; 2021. p. 1–10.
Leerssen P. The soap box as a black box: regulating transparency in social media recommender systems. Eur J Law Technol 2020;11(2).
Lulamae J. How the "shadow banning" mystery is messing with climate activists' heads. Algorithm Watch; 2022 Available at: https://perma.cc/JYC5-BQ82 (accessed 28 June 2022).
Mahieu RL, Ausloos J. Harnessing the collective potential of GDPR access rights: towards an ecology of transparency. Int Policy Rev 2020. Available at: https://policyreview.info/articles/news/harnessing-collective-potential-gdpr-access-rights-towards-ecology-transparency/1487.
Merrill J, Oremus W. Five points for anger, one for a 'like': how Facebook's formula fostered rage and misinformation. Washington Post 2021. 26 October. Available at: https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/(accessed 4 January 2023) [https://perma.cc/7UQR-JQES].
Miller EL. Amplified Speech. Cardozo Law Rev 2021;43:1.
Myers West S. Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. New Media Soc 2018;20(11).
Napoli PM. Social media and the public interest: governance of news platforms in the realm of individual and algorithmic gatekeepers. Telecomm Policy 2015;39(9):751–60.
Nicholas G. Shadowbanning is big tech's big problem. The Atlantic; 2022 April 28. Available at: https://www.theatlantic.com/technology/archive/2022/04/social-media-shadowbans-tiktok-twitter/629702/ (accessed 28 June 2022).
Pasquale F. The Black Box Society: The secret Algorithms That Control Money and Information. Harvard University Press; 2015.
Poell T, Nieborg D, Duffy B. Platforms and Cultural Production. John Wiley & Sons; 2022.
Radsch C. Shadowban /shadow banning. In: Belli Luca, Zingales Nicolo, Curzi Yasmin, editors. IGF Glossary of Platform Law and Policy Terms. Internet Governance Forum; 2021 Available at: https://perma.cc/7K3Q-HN36.
Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services.
Rieder B, Hofmann J. Towards platform observability. Int Policy Rev 2020;9(4):1–28.
Roberts ST. Behind the Screen. Yale University Press; 2019.
Suzor N. Digital constitutionalism: using the rule of law to evaluate the legitimacy of governance by platforms. Soc Media+ Soc 2018;4(3).
Selbst A, Powles J. Meaningful information and the right to explanation. Int Data Privacy Law 2017;7(4).
Seaver N. Algorithms as culture: some tactics for the ethnography of algorithmic systems. Big Data Soc 2017;4(2).
Suzor NP, West SM, Quodling A, York J. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation'. Int J Commun 2019;13:18.
Swisher K. Zuckerberg: the recode interview. Vox; 2018 Available at: https://perma.cc/45J9-CL2R (accessed 28 June 2022).
Thorburne L, Stray J, Bengagina P. What will "amplification" mean in court?. Tech Policy Press; 2022 Available at: https://techpolicy.press/what-will-amplification-mean-in-court/ (accessed 4 January 2022).
Thorson K, Wells C. Curated flows: a framework for mapping media exposure in the digital age. Commun Theory 2016;26(3):309–28.
Twitter. Setting the record straight on shadow banning. Twitter Blog; 2018 Available at: https://perma.cc/XTZ4-BDYH (accessed 28 June 2022).
Van Couvering E. Is relevance relevant? Market, science, and war: discourses of search engine quality. J Comput-Med Commun 2007;12(3):866–87.
Van Hoboken J. Search Engine freedom: On the Implications of the Right to Freedom of Expression For the Legal Governance of Web Search Engines. Kluwer Law International; 2012.
Waldron J. The rule of law. Stanford Encyclopedia of Philosophy; 2016. Available at: https://plato.stanford.edu/entries/rule-of-law/#ProcAspe(Accessed 28 June 2022) [https://perma.cc/P55L-57HZ].
Wisner M. Google's Eric Schmidt responds to verizon, AT&T pulling ads from YouTube. Fox Business; 2017. Available at: https://www.foxbusiness.com/features/googles-eric-schmidt-responds-to-verizon-att-pulling-ads-from-youtube(accessed 28 June 2022)[https://perma.cc/HW8S-A259].
Wodinsky S. YouTube's copyright strikes have become a tool for extortion. The Verge; 2019 11 February. Available at: https://www.theverge.com/2019/2/11/18220032/youtube-copystrike-blackmail-three-strikes-copyright-violation (accessed 4 January 2023).
Zalnierute M. Transparency washing" in the digital age: a corporate agenda of procedural fetishism. Crit Anal Law 2021;8(1).