Lia Eggleston

Professor Joan Feigenbaum

CPSC 610

December 17, 2021

Transparency in Content Moderation: Overview and the Case of "Shadowbans"

I. Introduction

There have long been calls to address the perceived lack of transparency in large-scale social media content moderation practices. There are various views on the role transparency should play, but it is generally considered an important goal for content moderation governance because it is viewed as a first step towards practical accountability for platforms, a means of making due process possible. The justification for this goal has a direct connection to legal processes, in which giving public reasoning behind decisions is seen as fundamental to accountable decision-making. [10] The legal context analogy is that "justice should not only be done, but should manifestly and undoubtedly be seen to be done". [19]

Opacity, or a lack of transparency, can be observed in the contexts of both operational methods and individual application of rules for moderation. [5] While transparency is often used as a catch-all term always leading to improvement, it can occur in many specific ways within a platform's practices. Many calls for transparency, however, are also "so vague as to not be useful" in linking transparency to accountability. [5]. Therefore, any solutions that claim to do so should aim for "specificity in identifying what information should be provided and to whom". [10] This analysis will address recommendations for such specificity in section V, Solution Spaces.

Section II describes the harms associated with opacity generally and in several contexts, including automated moderation and transparency reporting. Section III raises some of the challenges and caveats of addressing opacity. Section IV narrows in on a specific set of practices colloquially known as “shadowbanning”, its uses and observed harms. Section V analyzes solution spaces that have been proposed so far. Finally, I draw on the solution space to bring together some recommendations moving forward in Section VI.

II. Observed Harms of Opacity

A. General

A lack of information can make it difficult to have informed public debate about content moderation practices and the complex interests that go into them. When neither the user base, nor researchers or governments feel that they can understand the reasoning behind moderation practices, it is difficult to analyze what policies may be in the public interest or not.

In a 2019 study only 50 percent of users were confident that they understood the reasons for a moderation decision. [10] And while it may make sense to hide some of the criteria in order to make it harder for users posting harmful content to evade flagging, even “trusted third-party auditors and vetted researchers” generally cannot access details about major moderation processes. [1]

Low transparency often leads to the confusion, frustration, and sometimes exclusion of users. Unable to know what factors affect decisions about their content, users are left to develop folk theories of reasons such as conspiracy and bias. [4] This can lead to self-censorship in an attempt to avoid moderation, or departure from platforms. [5] When moderation decisions are actually affected by systemic bias, intended or not, opacity about the decisions can also hide the harm of that bias. For example, in 2017 YouTube was criticized for demonetizing LGBTQ educational videos and hiding them under "restricted mode", but attempted to evade blame due to vague information about their policies. [10] In addition, lack of transparency can tie in to the problem of labor abuses for human content moderators, because companies are loath to disclose exactly what amount of work is done and how, by human moderators vs. automation. [8] In 2017 as well, Facebook pledged to add 3000 moderators to their workforce to address concerns about capacity but also gave no indication of who they would hire, where, and under what working conditions. [8]

B. Algorithmic Content Moderation

Specifically, algorithmic content moderation can produce a Black Box effect that exacerbates the effects of opacity for moderation rules and processes, largely because it becomes much more complex to determine what exactly led to a flagging or takedown decision. [1] Users often don't know whether or to what degree automation was used in a decision, or any of the rules and values that may have been built into a moderation system by its designers. For large platforms, this can also create an image of objective technology as the ultimate solution in an attempt to de-politicize moderation outcomes and avoid accountability. [8] In addition, fraught cases of wrong or deceptive output can lead to even more user distrust and confusion than with human moderators. [9] This can be seen in examples such as a Tumblr anti-porn algorithm flagging many non-sexual posts, or a Yelp review filtering that was accused of harming smaller-scale sellers on their site. [9]

C. Operational Rules and Transparency Reports

Above the individual level, standards for decisional transparency in operational choice rules are also lacking. Major platforms do currently release “transparency reports” regularly, but there aren’t formal standards for these and they usually “report only takedown requests from governments and companies.” [5] Sometimes these also give aggregate numbers on large categories such as takedowns of terrorist material, but those aggregations are often insufficient to analyze how moderation systems are actually working or how they could be improved. [10] For example, Facebook in 2019 claimed to have removed 99.6 percent of terrorist propaganda. [1] However, this avoids releasing answers to import inquiries such as what training data was used, what content or groups were considered to be terrorists, or any analysis of technical or human errors in the process. [1]

III. Challenges and Caveats of Improving Transparency

Even though the main goal of transparency may be to improve decision-making accountability for platforms in the future, “the extent to which transparency actually leads to greater accountability and better outcomes, however, is often unclear at best.” [10] Even when greater transparency can be achieved on high level operation rules, it can be a big challenge translating those decisions to individual applications of rules and it is not always possible to do so - for example linking an open decision made by the Facebook Oversight board to an individual content removal based on that open decision’s rule change. [2]

Also, the counter argument against greater transparency of moderation processes and reasoning is that too much exposure of process could lead to greater abuses on platforms by bad actors. [9] While this is important to keep in mind, the outcomes depend on the context. Research has supported both the idea that explaining moderation decisions is effective in reducing future harmful material, but also that users are persistent in finding ways around word filters when they know how the filters work. [11]

IV. Focus: Visibility Reduction and "Shadowbans"

A. Introduction to Visibility Reduction

Further insight can be gained through analysis of a specific platform practice that content moderation opacity brings to mind: the use of visibility reduction or "shadowbanning". The term "shadowban" is colloquial rather than technical, referring to a category of hiding or reduction in visibility of a user's content, without informing the user at all. This set of practices is sometimes used to handle content that a platform categorizes as "borderline" under their policies and so is not subject to full removal; or for any other case where the platform doesn't want the user to explicitly know about the action taken. [13]

Under what is most commonly known as a shadowban, a user's account may be kept active, but only the user can see their own content. [13] However, there are several types and scales of visibility reduction mechanisms, including lowering search term visibility, removal from recommendation, or reduction in distribution. Some examples of this include removal of content classified as "borderline" from being recommended by Youtube [13], or Facebook 'downranking' in which a piece of content is distributed less often but the poster is not notified like they would be for content removal. [3]

B. Uses and Legal Cases

One of the main reasons given for using shadowban-like practices is that it could make things more difficult for users who post harmful content, or especially spammers who use bots to post harmful content. In those cases, the idea is that spammers may not know when they've been detected and could delay or prevent bad actors from making new accounts and continuing. [10] For example, Facebook downranking is intentional and Facebook argues that posters of ‘borderline’ harmful content would have “fewer incentives to the commenting user to spam the page or attempt to circumvent the social networking system filters.” [3]

In addition, there are cases where it is legally required to withhold notifying users of content moderation actions. When a request comes under seal from law enforcement, such as in a court order or subpoena, the platform can be prohibited from notifying the user of the content moderation - but often only in cases where notification is deemed to pose a threat to a person or the public. [10] For example, secrecy can be compelled in court orders for national security cases such as for anti-terrorism, cybercrime investigation or espionage purposes. [20]

C. Observed Harms

By nature of both their perceived and confirmed existence on various platforms, shadowban practices can exacerbate user uncertainty and cause some groups to feel as though they are “shouting into a void”. [4] When there are a proliferation of unclear or excessive bans on some users’ accounts or content, a “threat of invisibility” effect can then cause or enhance user self-censorship. [5] In addition, when lack of transparency is already the norm for platforms, “Black Box Gaslighting” can occur when companies leverage their high opacity of moderation in order to “destabilize credible criticism” and be seen as the only authority on the effects of their algorithms. [14]. A case study of instagram users found that the users would start to believe company reassurances about visibility reductions even when contradictory evidence was found. [14] In other words, shadowban-like practices can further deflect accountability for harms.

Finally, some groups of users affected by shadowbanning of their posts of content such as body positivity, activism, or queer content were left without explanation of what affected their content visibility, but saw trends of double standards. [17] Users cannot see what logic marks their content into ‘borderline’ categories, but intentional or unintentional bias can easily be concealed when users aren’t even notified about the moderation action. This is especially important because visibility reduction practices can have real economic implications for some categories of users with livelihoods closely tied to platforms, such as small business owners, sex workes, or influencers. [17]

V. Solution Spaces

A. General

At its most basic, the way to address the harms discussed above is for platforms to provide more information, but also to do so in a productive way that could eventually promote accountability. This section analyzes various past proposals and recommendations regarding what information should be provided and how, in order to productively address opacity in moderation.

For individual moderation decisions or flagging, just providing the user a statement of which content was moderated and why could go a long way towards improving moderation; for example simple reason statements were shown to reduce the prevalence of harmful content on Reddit later on. [12] More specifically, Suzor et. al. recommend that the “URL of the prohibited content or a sufficiently detailed extract is available in the notification” and that notifications of moderation “should be permanently available to the user in some form” rather than cryptic or ephemeral. [10] Horten argues that individual notifications should include not only the basis of the decision but also “the process by which it was made.” [16] This is backed up by the recommendations of Naher et. al. around algorithmic content moderation, to give users information about both the existence and extent of algorithmic moderation in decision processes, even if the reasons for specific ML decisions are still difficult to backtrace fully. [9]

In order to improve the effectiveness of transparency reports, Suzor et. al. also emphasize the importance of providing data to researchers, recommending that the reports “provide the data that will enable researchers to understand a wide variety of concerns about moderation systems. This is the difference between transparency that merely provides aggregate information and transparency that can help to foster accountability.” [10] Various interactive interfaces for moderation have also been put forward as innovative ways to improve transparency, among other things. For example, PolicyKit software aims to promote community governance, and assumes accessibility to the full process and rules in use. [21] Naher et. al. propose a discussion interface prototype to improve user understanding and trust by letting users explore the moderation algorithm through text-based interactions. [9] Vaidya et. al. outline a visual analytic system that could help identify and maintain rules, as well as encourage detailed reasons for moderation decisions in a “human-in-the-loop” moderation system. [15]

B. Legal

There have been relatively few attempts to legislate around transparency for content moderation so far, but one interesting example to analyze is the German NetzDG or Network Enforcement Act. While primarily put forward in order to require combat of online hate speech and interesting for that reason too, NetzDG also imposes transparency requirements on platforms. Specifically, if a social media platform is at a scale to receive over 100 complaints per year, “it is required to publish semi-annual reports detailing its content moderation practices.” [18] That is, along with raw numbers of complaints and takedowns the transparency reports of companies operating in Germany also reported on moderating procedures. This aspect of the law has received “almost universal support” as well, despite high controversy for a lot of its other parts. [18] The Transatlantic Working Group [for Content Moderation Online] additionally recommends improvement on the NetzDG transparency requirement through measures like standardizing the report format, or ensuring more granular reporting data is available for researchers (something Suzor would agree with). [18] Towards that last end, the Working Group proposes the creation of government and platform-supported "research repositories that would combine data from multiple platforms". [18]

VI. Recommendations

I generally agree with Horten that on an individual level, users should almost always be notified of any moderation action on their content, and the notification should include both a reason and basic components of the moderation process. [16] In the specific case of shadow-ban like practices, the simple existence of those policies at all causes many of the harms discussed, so I would recommend ceasing those practices, and instead providing notifications very similar to those for removals, if any visibility reduction measures are taken. The only exceptions to this should be for legally required cases outlined in Section IV B and perhaps for clear bot spammers only; but beyond that the harm that can be caused by arbitrary or biased 'borderline' content shadowbans likely outweighs any benefits.

Second, standards for transparency reports would be very helpful, and I would join Suzor et. al. [10] and the Transatlantic Working Group [18] in recommending that reports facilitate research into more than top-level aggregate numbers.   Prioritizing research in areas like analyzing platform removal decisions and tracing through the black box effect of algorithmic moderation would also help to improve and utilize transparency measures effectively.

Finally, while I don't know how likely such legislation may be outside of the EU soon, I would recommend legislation similar to the transparency report requirement of NetzDG. The degree of specificity required for platforms to describe their moderation processes should be a subject of debate, but at the very least the extent of automation, rule creation process, and conditions of human moderators should be fair game to report.

Improving transparency is still a vital goal in content moderation governance, as it has the potential to open the way to accountability for large-scale platforms. All that is required to start the process towards improvement is attention to a few simple practices above, whether enforced or standardized.

Bibliography

[1] Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): 205395171989794. https://doi.org/10.1177/2053951719897945

[2] Klonick, Kate. 2021. “Inside the Making of Facebook’s Supreme Court.” The New Yorker. February 12, 2021. https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court

[3] Douek, Evelyn. 2019. “FACEBOOK’S ‘OVERSIGHT BOARD:’ MOVE FAST with STABLE INFRASTRUCTURE and HUMILITY.” North Carolina Journal of Law & Technology. https://ncjolt.org/wp-content/uploads/sites/4/2019/10/DouekIssue1_Final_.pdf.

[4] Myers West, Sarah. 2018. “Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms.” *New Media & Society* 20 (11): 4366–83. https://doi.org/10.1177/1461444818773059

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

[6] Suzor, Nicolas P. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives*. *Cambridge University Press*. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/lawless/8504E4EC8A74E539D701A04D3EE8D8DE

[7] “Santa Clara Principles on Transparency and Accountability in Content Moderation.” 2018. Santa Clara Principles. 2018. https://santaclaraprinciples.org/

[8] Roberts, Sarah T. 2018. “Digital Detritus: ‘Error’ and the Logic of Opacity in Social Media Content Moderation.” *First Monday* 23 (3). https://doi.org/10.5210/fm.v23i3.8283

[9] Naher, Jibon, Juho Kim, and Taehyeon An. n.d. “Improving Users’ Algorithmic Understandability and Trust in Content Moderation.” Accessed December 18, 2021. https://kixlab.github.io/website-files/2019/cscw2019-workshop-ContestabilityDesign-paper.pdf

[10] Suzor, Nicolas P., Sarah Myers West, Andrew Quodling, and Jillian York. 2019. “What Do We Mean When We Talk about Transparency? Toward Meaningful Transparency in

Commercial Content Moderation." *International Journal of Communication* 13 (0): 18. https://ijoc.org/index.php/ijoc/article/view/9736/2610

[11] Seering, Joseph. 2020. "Reconsidering Self-Moderation." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2): 1–28. https://doi.org/10.1145/3415178

[12] Jhaver, Shagun, Amy Bruckman, and Eric Gilbert. 2019. "Does Transparency in Moderation Really Matter?" *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–27. https://doi.org/10.1145/3359252

[13] Goldman, Eric. 2021. "Content Moderation Remedies." Papers.ssrn.com. Rochester, NY. 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3810580

[14] Cotter, Kelley. 2021. "'Shadowbanning Is Not a Thing': Black Box Gaslighting and the Power to Independently Know and Credibly Critique Algorithms." *Information, Communication & Society*, October, 1–18. https://doi.org/10.1080/1369118x.2021.1994624

[15] Vaidya, Sahaj, Jie Cai, Soumyadeep Basu, Azadeh Naderi, Donghee Yvette Wohn, and Aritra Dasgupta. 2021. "Conceptualizing Visual Analytic Interventions for Content Moderation." IEEE Xplore. October 1, 2021. https://doi.org/10.1109/VIS49827.2021.9623288

[16] Horten, Monica. 2021. "Algorithms Patrolling Content: Where's the Harm?" Papers.ssrn.com. Rochester, NY. February 22, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792097

[17] Dinar, Christina. 2021. "The State of Content Moderation for the LGBTIQA+ Community and the Role of the EU Digital Services Act." *Boell.org*. Heinrich-Böll-Stiftung. https://us.boell.org/sites/default/files/importedFiles/2021/06/21/HBS-e-paper-state-platform-moderation-for-LGBTQI-200621_FINAL.pdf

[18] Tworek, Heidi, and Paddy Leerssen. 2019. “An Analysis of Germany’s NetzDG Law.” *Annenberg Public Policy Center*. Transatlantic Working Group. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/NetzDG_TWG_Tworek_April_2019.pdf

[19] R v. Sussex Justices, Ex parte McCarthy, 1 KB 256 (High Court of Justice 1924).

[20] Woolery, Liz, Ryan Hal Budish, and Kevin Bankston. 2016. “The Transparency Reporting Toolkit: Best Practices for Reporting on U.S. Government Requests for User Information.” *Berkman Klein Center*. https://dash.harvard.edu/handle/1/28552578

[21] Amy Zhang, Grant Hugh, and Michael Bernstein. "PolicyKit: Building Governance in Online Communities"