Marginalization through Content Moderation:

An Interdisciplinary Analysis of the Silencing of Marginalized Voices on Social Media

Poverty and Human Capability Studies Capstone

Author: Katie Yurechko

Adviser: Professor Goldsmith

Word Count: 9,732

Introduction

Social media is largely considered a democratic space. It seemingly promotes the freedom and equality of all people by enabling individuals to voice their perspectives online, regardless of their identities. However, while many Americans use social media to share their opinions and experiences, algorithms disproportionately remove the content of users who hold marginalized identities without justification. Such unwarranted removals occur through the process of content moderation, whereby social media platforms attempt to rid the digital realm of content that violates their rules, including violent content and harassment. Cloaked in opacity, the silencing of marginalized users via content moderation not only reflects ongoing discrimination against marginalized groups, but it reinscribes systemic prejudices like racism and sexism into the physical world, often without public awareness.

Throughout my paper, I attempt to answer the following question: *What, if anything, should be done regarding social media content moderation policies and their enforcement to address the problems that they aim to address but without disadvantages for marginalized groups*? The threat that unfairness in content moderation poses to marginalized communities makes formulating and pursuing a response to this question into a moral imperative. Considering the silencing of marginalized individuals from a Rawlsian perspective, one understands that behind a veil of ignorance that prevents us from knowing our identity and personal circumstances, we would devise principles of justice that are not biased for or against any specific group. In particular, if someone did not know which racial or gender identity they would bear, they would choose to structure content moderation such that it would not harm the marginalized groups to which they could belong. This Rawlsian conception of justice as fairness enables us to acknowledge that we would not allow the silencing of voices that results in an inability of one's interests to be heard, and therefore increases the potential for others to ignore or violate one's interests, if we were ignorant of our own identities.[1] It is thus clear that Rawlsian justice requires us to intervene in unjust content moderation practices to reduce the silencing of individuals with marginalized identities online.

I structure my paper in three parts:

---

[1] Rawls, John. From *A Theory of Justice*. In *Ethical Theory and Business*. 8th Edition. Ed. Tom L. Beauchamp. Pearson/Prentice Hall, 2009.

i) *Literature Review*: I describe social media content moderation, then explain the impacts of content moderation on marginalized groups as well as potential solutions to the negative impacts.
ii) *Methodology*: I detail the approach that I will use to address my research question.
iii) *Analysis*: I evaluate the potential solutions that I identified in my *Literature Review* and ultimately propose my own solution to the disproportionate silencing of marginalized voices on social media.

Finally, the *Conclusion* summarizes my paper, providing a holistic overview of the unfairness in content moderation in addition to my proposed solution to the problem.

Literature Review

To analyze and address issues in social media content moderation, particularly the disproportionate silencing of marginalized voices on social media platforms, one must first understand how content moderation on social media operates and impacts marginalized communities. This section in particular aims to examine three topics. First, I explain existing social media content moderation policies and enforcement mechanisms, which involve human content moderators, users, and algorithms. Second, I discuss the impacts of content moderation on marginalized communities as well as the causes behind these impacts. Finally, I describe the efforts that have been made with the potential to address content moderation biases. I explore each of these topics in the three subsections below.

Content Moderation Policies and Enforcement Mechanisms

Social media companies like Meta, which governs Facebook and Instagram, espouse missions of "giving people the power to build community and bring the world closer together."[2] Twitter and TikTok share similar goals "to give everyone the power to create and share ideas and information"[3] and "to inspire creativity and bring joy,"[4] respectively. To facilitate their goals of fostering community-building and creative expression, these companies determine and enforce community guidelines that serve as standards of behavior on their platforms. The community guidelines and their enforcement mechanisms comprise the process of content moderation, which

[2] https://about.meta.com/company-info/
[3] https://investor.twitterinc.com/contact/faq/default.aspx
[4] https://www.tiktok.com/about?lang=en

is "the organized practice of screening user-generated content (UGC)" to facilitate safe environments online.[5]

I. Policies

Community guidelines are publicly accessible policies that outline problematic activities like hate speech, violence, and misinformation on social media platforms. A study by the Partnership for Countering Influence Operations (PCIO)[i] compiled and analyzed the community guidelines of thirteen social media and messaging platforms, including Facebook, Instagram, TikTok, and Twitter. The study found that while some platforms employ more generalized approaches, using vague language to describe prohibited activity, other platforms implement detailed policies that describe specific violations. Community guidelines additionally differ in their length and complexity, with Twitter's guidelines totaling 23,110 words and detailing 29 policies while Instagram's guidelines only total 208 words and detail 2 policies. *Generalized Policies* give platforms the flexibility to tweak and interpret their rules as they choose, whereas *Particularized Policies* are more inflexible but allow for greater transparency to users regarding what content is prohibited.[6ii]

Beyond terminology, the substance of community guidelines differs by platform. Most platforms ban harassment and threats, spam, and violent content, but they tend to disagree about how to respond to child sexual abuse material, false information, and hate speech.[7] For example, while TikTok clearly describes its understanding of and attempt to regulate hate speech ("Hateful ideologies are those that demonstrate clear hostility toward people because of their protected attributes. Hateful ideologies are incompatible with the inclusive and supportive community that our platform provides and we remove content that promotes them"[8]), Instagram simply writes, "We remove content that contains credible threats or hate speech"[9] without elaborating further.[iii]

II. Enforcement Mechanisms

The enforcement of community guidelines often relies on three mechanisms: human users, human content moderators, and algorithmic tools. Users of social media platforms flag or

---

[5]https://journals.sagepub.com/doi/pdf/10.1177/1461444818773059?casa_token=z3LosI5N9IAAAAAA:VcbKUN8bqh-Pk96M0DM7MysMyAdSTa6Oh0rCqThqaRNZLtOqejTtMfpPKZ_ovq-WM104hoVOguW8bw

[6] https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201

[7] https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201

[8] https://www.tiktok.com/community-guidelines?lang=en#38

[9] https://help.instagram.com/477434105621119

report content that they consider objectionable, which provides platforms with hand-selected, potentially inappropriate content for their human content moderators to review. These workers scroll minute-by-minute through that content and more, manually determining which content violates their platform's community guidelines. Finally, platforms utilize artificial intelligence (AI) for content screening.[10] For example, AI-backed content moderation can automatically review textual, audial, and visual content for community guidelines violations.[11]

Inappropriate content can be identified during the upload process (i.e., when a user chooses to click "post") or at any point after it is uploaded to a platform. Once inappropriate content is identified, users can experience a range of consequences, including content deletions, account bans, and shadow-banning, where content is made invisible to other users without actually being deleted from the platform. Users report that hindered content visibility or the loss of platform access not only hampers their capacity for self-expression; it impedes involvement in online communities and professional networks that underly social, political, and economic life. Though platforms allow users to appeal account bans and content removals that users believe were doled out in error, most users express that content moderation systems do not give them sufficient opportunity to channel their agency into challenging the consequences that had been assigned to them.[9]

Understandings of Content Moderation Impacts on Marginalized Communities

In this paper, I define marginalized communities to be groups that experience social, political, or economic discrimination or exclusion due to historically unequal power dynamics within society. In this section, I explain the impacts of content moderation on marginalized communities and detail potential causes of these impacts. Though I will predominantly discuss the Black and LGBTQ+ (Lesbian, Gay, Bisexual, Transgender, and Queer) communities throughout this section, I acknowledge the diversity of communities that experience marginalization and the challenges that they might encounter in the digital realm.

I. Impacts

---

[10] https://journals.sagepub.com/doi/pdf/10.1177/1461444818773059?casa_token=z3LosI5N9IAAAAAA:VcbKUN8bqh-Pk96M0DM7MysMyAdSTa6Oh0rCqThqaRNZLtOqejTtMfpPKZ_ovq-WM104hoVOguW8bw

[11] https://www.forbes.com/sites/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation/?sh=f7a230d4a178

The #MoveMe guide to social movements and social media created by UC Berkeley outlines twenty-six movements that social media amplified by raising attention in the public consciousness. The paradigm example of social media facilitating social organization is the *Arab Spring*, also known as the *Arab Revolutions*, which utilized social media to spread information about pro-democracy efforts and gain support from a global audience. The ongoing *Black Lives Matter* (*BLM*) and *LGBTQ+ Rights* movements also reap similar benefits from social media. For example, BLM was catalyzed by Twitter in 2013 and gained greater momentum when a video of George Floyd being suffocated by police officers circulated social media platforms, forcing viewers to confront the realities of systematic racism that overwhelmingly impact the Black community in the United States. Today, the number of social media posts using the hashtag #BlackLivesMatter totals over forty million, which demonstrates the digital scope of the movement. In addition, the LGBTQ+ community gains awareness and visibility through social media, while its members learn about the community, find guidance, and share their lived experiences. Facebook in particular implemented options for LGBTQ+ support, allowing users to utilize the "rainbow flag reaction" to express support for LGBTQ+ media during Pride Month.[12]

While social media strengthens the social movements of marginalized communities, with past U.S. Supreme Court Justice Anthony Kennedy stating that social media allows someone with Internet connectivity to "become a town crier with a voice that resonates farther than it could from any soapbox,"[13] research shows that social media often silences or excludes marginalized users. For example, a 2021 news report revealed that Twitter's image-cropping algorithm would reliably crop out Black people in pictures of Black and white people. For example, with photos that included Barack Obama and Mitch McConnell, the algorithm cropped the pictures to show only Mitch McConnell. Twitter then apologized and discontinued its image-cropping algorithm.[14]

The majority of issues regarding the suppression of marginalized communities on social media involves content moderation. For example, The Washington Post references Facebook

[12] https://moveme.berkeley.edu/project/lgbtqrights/
[13] https://www.forbes.com/sites/kalevleetaru/2017/10/14/how-social-media-can-silence-instead-of-empower/?sh=7038d1c27ba1
[14] https://www.nbcnews.com/tech/tech-news/twitters-racist-algorithm-also-ageist-ableist-islamaphobic-researchers-rcna1632

documents that detail the negative impacts of its previous race-blind model on Black users. The model regarded hate speech against all people equally, without considerations of race, which led "statements of contempt, inferiority, and disgust directed at White people and men" to comprise 90 percent of all hate speech that was removed by Facebook in 2020. [15] The Facebook algorithm has since been revised to prioritize the removal of hate speech against minorities.[16] In addition, a Media Matters for America study in 2020 uncovered a bias against positive or neutral transgender content in favor of anti-trans posts on Facebook. The study found that 43% of posts about trans issues that earned at least 50,000 interactions involved negative attitudes toward trans athletes, while only 11% were about health care for trans youth.[17] These examples show that social media algorithms perpetuate discrimination against the Black and LGBTQ+ communities.[iv]

II. Causes of Negative Impacts

Researchers provide three predominant explanations for the unjustified deletion of content related to marginalized communities. First, even though users report content that they find objectionable and human content moderators review it, the large volume of reports received each day as well as language subtleties (e.g., marginalized communities reclaiming words like the N-word) often prevent the *accurate* moderation of content, meaning that content which fails to violate a platform's community guidelines nonetheless faces removal from the platform. In addition, reporting is often driven by partisanship and personal ideology, which can promote the elimination of unpopular speech. Second, platforms' increasing reliance on AI for the identification of rule-violating content can automate biases in the content moderation process. For example, computational linguists found that tweets written in African-American Vernacular English (AAVE) commonly spoken by Black Americans are nearly twice as likely to be flagged as offensive compared to others. The insensitivity of content annotators to dialectical differences promoted this racial bias in automatic hate speech detection models, which were trained on annotator data.[18] In summary, although natural language processing is often perceived as an objective tool to identify rule-violating content, algorithmic systems can misclassify content

[15] https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/
[16] https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/
[17] https://www.mediamatters.org/facebook/right-leaning-facebook-pages-earned-nearly-two-thirds-interactions-posts-about-trans
[18] https://aclanthology.org/P19-1163.pdf

based on contextual details that might not be understood by the human beings who train the algorithms. Third, some researchers explain biases by pointing to platforms' business models. For example, an AI ethicist describes, "Facebook profits from the proliferation of extremism, bullying, hate speech, disinformation, conspiracy theory, and rhetorical violence." This is because more extreme content attracts users to spend more time engaging with a platform, thus raising platform revenues through more targeted advertising. For this reason, the ethicist concludes that "Facebook's problem is not a technology problem. It is a business model problem."[19]

Solutions to Social Media Biases

A multiplicity of perspectives regarding the causes of content moderation biases produce a variety of advised solutions, which I divide into four categories: *fix the technology*, *reimagine the technology*, *fix the system at large*, and *reimagine the system at large*.

I. Fix the Technology

Technology teams often seek to remedy existing content moderation technologies, which involves tweaking the algorithmic processes that identify inappropriate content. For example, Stanford's Artificial Intelligence Laboratory is working to develop a more inclusive content moderation algorithm for platforms like Reddit and Discord. The researchers shared, "We wanted to empower the people deploying machine learning models to make explicit choices about which voices their models reflect." They explain that the Stanford algorithm is a "jury learning algorithm," which simulates individual annotators whose data trains machine learning models. By modeling annotators, jury learning allows social media executives to mix and match different identities to comprise a fair and representative jury for evaluating content. The Stanford team also plans to build an ethical framework to guide executives in selecting a representative jury to govern the algorithm. [20] In addition, European researchers champion a "human-is-the-loop" approach to semi-automated content moderation. Unlike the typical "human-in-the-loop" approach whereby algorithms at times consult human beings to judge the accuracy of their content moderation, the "human-is-the-loop" approach gives human content moderators ultimate control in deciding which content should or should not be removed from a platform.[21]

---

[19] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8179701/
[20] https://stanforddaily.com/2022/10/04/stanford-ai-content-moderation/
[21] http://idl.iscram.org/files/daniellink/2016/1401_DanielLink_etal2016.pdf

II. Reimagine the Technology

Some technologists take approaches besides improving existing content moderation algorithms, instead seeking to reimagine the standard approach to content moderation. For example, MIT researchers developed an experimental platform that gave users control over content moderation, which showed that users do evaluate content effectively and collaboratively.[22] Their study proposed leveraging trusted-peer assessments to combat misinformation online, rather than rectifying the algorithms that already exist. The trusted-peer assessments would entail having each user indicate their trust in a subset of other users with whom they interact online; the user's trusted connections would then rate the truthfulness of the user's content when it is posted, thus enabling the content to be removed from the platform based on the cumulative judgement of trusted online connections. [23] For example, if I am a trusted connection of a user who posts that the Earth is flat, I would give the user a very low accuracy rating, which could result in the removal of the user's post depending on the assessments of their other trusted connections. This approach bears similarity to the process already adopted by Reddit. The platform uses community leader-moderators who enforce clear rules about what is prohibited from the platform and how violators will be punished. Each subreddit (i.e., micro-community within the Reddit platform that discusses a particular topic) operates under different rules, since leader-moderators specify rules that are relevant to the subreddit's topical area. For example, the leader-moderator of a subreddit that focuses on video games might specify which forms of violence are acceptable (e.g., digital explosions) and which forms are unacceptable (e.g., digital murders). Leader-moderators then remove content that they determine to be in violation of the subreddit's guidelines.[24] As a result of this user-led process, volunteer content moderators on Reddit save the company $3.4 million worth of work every year.[25]

III. Fix the System at Large

Numerous groups, including national governments, tend to favor fixing the systems (i.e., social media company contexts) in which content moderation operates. Such remedies often involve increasing transparency into social media companies and holding the companies

---

[22] https://news.mit.edu/2022/social-media-users-assess-content-1116
[23] https://dl.acm.org/doi/10.1145/3555637
[24] https://hbr.org/2022/11/content-moderation-is-terrible-by-design
[25] https://www.newscientist.com/article/2325828-reddit-moderators-do-3-4-million-worth-of-unpaid-work-each-year/

accountable for their harms through fines and other legal ramifications, like payments to the victims of their harms.[v] For instance, various organizations and academics developed the Santa Clara Principles on Transparency and Accountability Around Content Moderation in 2018 to outline standards that social media platforms should meet regarding content moderation policies and their enforcement. The Principles ask that platforms publish "clear and precise rules and policies" as well as "ensure that their content moderation systems, including both automated and non-automated components, work reliably and effectively." However, though platforms like Facebook and Twitter have committed to adhering to the Principles, very few companies have fully met the outlined demands.[26]

Politically, there have been hundreds of failed attempts to regulate social media platforms in the U.S. Although senators have introduced policies like the *Platform Accountability and Transparency Act* aimed at giving researchers access to social media data and providing public transparency into content moderation, many of these acts have not received enough political consensus due to industry lobbying from social media companies.[27][28] Amy Klobuchar and other congresspeople who aim to remedy social media continue to propose acts like the *Social Media NUDGE Act*, which would require researchers to conduct ongoing studies aimed at the reduction of algorithmic harms.[29] Meanwhile, the European Union successfully passed the *Digital Services Act* in late 2022 to increase transparency in online platforms, give users more agency in content moderation through the appeals process, and provide researchers with access to platform data.[30]

IV. Reimagine the System at Large

Some ethicists and technology experts suggest that we should stop looking to technology to solve content moderations' flaws. More than that, we should abandon trying to remedy social media companies' operations and instead recognize the irredeemable flaws in the companies' current structures. For example, researcher Tarleton Gillespie explains, "The persistent failure of social media platforms to build moderation architectures that work should tell us something. Perhaps, if moderation is so overwhelming at this scale, it should be understood as a limiting

---

[26] https://santaclaraprinciples.org
[27] https://www.humanetech.com/insights/policy-brief-state-of-global-tech-policy
[28] https://techpolicy.press/the-platform-accountability-and-transparency-act-take-two/
[29] https://www.brookings.edu/blog/techtank/2022/02/23/senator-klobuchar-nudges-social-media-companies-to-improve-content-moderation/
[30] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en

factor on the 'growth at all costs' mentality. Maybe some platforms are simply too big."[31] Antitrust regulators agree with this conviction, arguing that competition will force companies like Facebook to fix their problems with content moderation.[32]

Methodology

To address my core research question (*What, if anything, should be done regarding social media content moderation policies and their enforcement to address the problems that they aim to address but without disadvantages for marginalized groups?*), I will first evaluate the potential solutions that I identified in my literature review. I will analyze the *practical effectiveness*, *moral permissibility*, and *political feasibility* of the four categories of solutions that I described (i.e., *fix the technology*, *reimagine the technology*, *fix the system at large*, and *reimagine the system at large*) by synthesizing literature across the disciplines of computer science, philosophy, and politics and interpreting the literature from the perspectives of the American government, social media companies, and social media users.

Second, I will compare my evaluations with one another to arrive at my proposed solution to the disproportionate silencing of marginalized groups via content moderation. I will describe and evaluate my solution for its *practical effectiveness*, *moral permissibility*, and *political feasibility* by comparing it to the four categories of solutions that I already inspected. To determine whether my solution would be *practically effective* at remedying content moderation issues, I will record insights from Dr. Simon Levy, an artificial intelligence and computational linguistics expert. To judge its *moral permissibility*, I will discuss my solution with Dr. Nathaniel Goldberg and Dr. Howard Pickett, both skilled ethical thinkers.[vi] These combined technical and philosophical perspectives will enable me to evaluate my solution in a holistic manner. Finally, to gauge whether my solution is *politically feasible*, I will consult policy briefs and other political documents that will provide insights into the practicality of my solution.

Analysis

Before conducting my analysis, I will briefly summarize the four categories of solutions to the silencing of marginalized voices on social media that I described above. The first category is *fix the technology*, which involves improving the algorithmic processes that identify inappropriate content. The second category is *reimagine the technology*, which seeks not to

[31] https://journals.sagepub.com/doi/full/10.1177/2053951720943234
[32] https://hbr.org/2020/09/breaking-up-facebook-wont-fix-social-media

improve existing algorithms but to change the standard approach to content moderation. The third category is *fix the system at large*, which involves shifting the contexts in which social media companies operate, including by combatting the opacity in their content moderation practices. The fourth category is *reimagine the system at large*, which involves acknowledging irreconcilable flaws in social media companies' structures and attempting to remedy them.

Now, I will analyze the *i) practical effectiveness*, *ii) moral permissibility*, and *iii) political feasibility* of these four categories of solutions. For each category, I will assign a ranking of low, moderate, or high to indicate the extent to which each category exhibits *i)*, *ii)*, and *iii)* in regards to the silencing of marginalized voices via content moderation on social media. Finally, I will propose my own solution to the aforementioned issue with content moderation, justifying its practical effectiveness, moral permissibility, and political feasibility by comparing it to the four categories of solutions that I evaluate below.

Fix the technology

Two primary examples demonstrate the approach of fixing content moderation technologies. The first is the Stanford jury learning algorithm, which is trained by a fair and representative "jury" of human beings to moderate content. The second is the "human-is-the-loop" approach, which turns algorithms from primary agents in content moderation into mere assistants while human beings hold the predominant role in identifying rule-violating content. Though both examples seek to modify how content moderation algorithms operate, the first focuses on how human beings train machine learning models to make decisions on their own, while the second places ultimate control not with artificial intelligence but with human beings themselves. I will evaluate these examples in tandem to judge the practical effectiveness, moral permissibility, and political feasibility of fixing content moderation technologies.

i) Practical effectiveness (moderate)

According to Professor Levy in his Artificial Intelligence seminar, the success of a machine learning model largely depends on the data used to train the model. For example, if one wants to use a model to discriminate between pictures of cats and dogs, one must first give the model many images of cats labeled as cats and many images of dogs labeled as dogs. Then, after one trains the model on these labeled images, one can evaluate the accuracy of the model by running it on a set of test data, which contains pictures of cats and dogs that the model has not yet seen. The model is more likely to produce correct classifications of cats and dogs if the data

on which it was initially trained is representative of the images that it could feasibly encounter. Therefore, attempts like the Stanford jury learning approach appear promising at countering the unfair removal of content from marginalized communities by improving the representativeness of the data on which models are trained. For instance, cases in which content moderation algorithms incorrectly interpret the N-word as hate speech could be reduced by training models on more text clips and images of Black people using the N-word in reclaimed ways.

However, Professor Levy explains that other factors besides the data on which a model is trained are pivotal to the success of machine learning. He describes hyperparameters, which are adjustable values that guide a machine to learn data. By simply tweaking the values of a model's hyperparameters, one can get radically different results from a model. For example, increasing the number of times that an algorithm iterates through (i.e., observes) training data can improve the accuracy of a model at classifying new data. But for every problem that machine learning attempts to solve, different hyperparameter values are optimal, and it can be not only challenging but impossible to find values that eliminate errors in data classification. For example, Professor Levy detailed the well-known MNIST database (Modified National Institute of Standards and Technology database), which contains examples of handwritten digits that programmers commonly use to train and test machine learning models. Though the best known classifier of MNIST digits has an impressive accuracy rate of 99.8%, the 0.2% error rate can be catastrophic at a large scale. In the moderation of content on social media, with millions of posts being classified each day, a near-perfect accuracy rate can still enable many errors. Instagram, for example, has nearly 100 million new posts daily.[33] Even if the Instagram classification system is as strong as the best MNIST digit classifier, it still misclassifies 200,000 posts every day, meaning that it either allows problematic content to remain on the platform or wrongly removes acceptable content. Thus, while algorithmic improvements may reduce the misclassification of content, they are not sufficient solutions to the issue of misclassification in content moderation.

Adjusting how the algorithms operate in conjunction with human beings (e.g., by giving humans more control over content moderation through the "human-is-the-loop" approach) appears to be a promising solution amidst the limitations of purely algorithmic content moderation. However, as elucidated by the Stanford jury learning algorithm, a team of human content moderators must be representative in order to reduce the potential of biases impeding the

[33] https://www.zippia.com/advice/instagram-statistics/

fairness of content moderation. But even with a representative team of content moderators, predominantly human-led content moderation might nonetheless be ineffective. The enormous scale of social media platforms prevents a thorough human-driven evaluation of content, since it is impossible for human content moderators to review the millions of posts uploaded to a platform each day. Even if approaches like "human-is-the-loop" support human beings by providing algorithmic suggestions about what content to review, algorithmic suggestions themselves often fall prey to inaccuracy. Various other challenges, including the need to train a diverse group of workers to apply a single set of rules to posts that differ greatly and require contextual interpretations, impede the effectiveness of human content moderation.[34]

Therefore, the effectiveness of fixing the technology is *moderate* because of present limits on algorithmic optimization coupled with the challenges of human content moderation.

ii) Moral permissibility (moderate)

While there are no foreseeable ethical issues posed by improving content moderation algorithms, the seeming need for human beings to moderate content (due to the present limitations of artificial intelligence) presents a moral dilemma.

In 2019, The Verge reporter Casey Newton interviewed content moderators working for the third-party company Cognizant, which Facebook previously used to screen its content. Newton detailed the severe mental health issues that content moderators developed from their jobs, with one employee stating, "I don't think it's possible to do the job and not come out of it with some acute stress disorder or PTSD."[35] In 2020, former content moderators sued Facebook over PTSD and trauma, claiming that they had to view thousands of "videos, images and live-streamed broadcasts of child sexual abuse, rape, torture, bestiality, beheadings, suicide and murder" each day with insufficient protections from Facebook, which resulted in a class-action lawsuit with a settlement of $52 million.[36,37] Clearly, the job of content moderation directly threatens mental wellbeing.

But it is important to acknowledge that human content moderation prevents millions of users from viewing trauma-inducing content that artificial intelligence fails to remove. One

---

[34] https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[36] https://futurism.com/the-byte/facebook-content-moderators-lawsuit-ptsd

[37] https://www.npr.org/2020/05/12/854998616/in-settlement-facebook-to-pay-52-million-to-content-moderators-with-ptsd

content moderator interviewed by Newton succinctly captures the importance of human content moderators: "If we weren't there doing that job, Facebook would be so ugly."[2] A utilitarian perspective underscores the moral necessity of human content moderation, since a few thousand workers endure pain to maximize the benefits of a much larger mass of users. In addition, according to a deontological perspective, the moral nature of an action is determined in relation to the responsibilities of those who perform it. Deontology would not oppose human content moderation because it is predominantly used to provide social media users with a safer online experience, thus fulfilling its obligation to promote safe digital communities. Nonetheless, the exposure of human beings to immense psychological distress serves as a strong barrier to the moral permissibility of human content moderation.

Therefore, the moral permissibility of fixing the technology is *moderate* because although improving artificial intelligence is entirely morally permissible, the current need to involve human beings in content moderation (due to the inadequacy of purely automated content moderation) is ethically problematic.

iii) Political feasibility (low)

Despite bipartisan support for regulating social media, industry lobbying has delayed hundreds of U.S. proposals to regulate platforms.[38] In addition, evidence suggests that policymakers do not have enough access to algorithmic knowledge regarding social media to inform effective policies. According to previous Facebook employee Frances Haugen, only about 300 to 400 experts understand and determine how social media algorithms work.[39] Mark MacCarthy, a government expert at the Center for Technology Innovation, additionally explains, "The more policymakers know about the inner workings of social media companies through transparency disclosures, the more likely it is that they will be able to devise ways to improve content moderation."[40] Thus, an understanding of how the algorithms work seems essential to crafting policies that actually improve content moderation, yet policymakers lack access to such an understanding.

[38] https://www.humanetech.com/insights/policy-brief-state-of-global-tech-policy
[39] https://ide.mit.edu/insights/transparency-the-first-step-to-fixing-social-media/
[40] https://www.brookings.edu/blog/techtank/2022/11/01/transparency-is-essential-for-effective-social-media-regulation/

Therefore, the political feasibility of fixing the technology is *low* because of powerful industry opposition and a general lack of knowledge into social media algorithms in the policymaking world.

Reimagine the Technology

Two key examples of reimaging the technology involve placing more power over content moderation into the hands of platform users. The first example leverages trust in others to assess content, giving users the ability to delegate content moderation to trusted connections online. The second example, adopted by Reddit, entails having users voluntarily moderate content within topical areas. I will evaluate these examples in tandem to judge the practical effectiveness, moral permissibility, and political feasibility of reimagining content moderation technologies.

i) Practical effectiveness (moderate)

The researchers behind trusted-peer assessments of content explain that their user study reveals how "people actually do treat content with scrutiny and they also try to help each other" reduce the spread of misinformation.[41] Though the study focused on preventing the spread of misinformation, it suggests that peer-led content moderation can operate effectively at scale. Writers from the Nebraska Law Review specifically state that a decentralized system of content moderation that puts more power into the hands of users "allows most moderation decisions to be made by users who are intimately aware of the contours of individual online communities." They add that "[s]uch decisions are less likely to restrict content that is not violative of an online communities' standards."[42] This finding explicitly suggests that trusted-peer assessments can impede unjustified content removals. Content moderation systems could rely on users to provide structured assessments of posts, determining whether content contains misinformation, hate speech, and other topics that platforms seek to remove.

Presently, platforms like Facebook struggle with quickly determining the contextual meaning of content, since human content moderators whom the companies hire cannot feasibly review large amounts of content with accuracy and algorithms are often unable to discern the context of words. For example, it is difficult for algorithms to judge whether a post that describes a police shooting is simply documenting the event or is praising or condemning it, which might

[41] https://news.mit.edu/2022/social-media-users-assess-content-1116
[42] https://lawreview.unl.edu/federalists-internet-what-online-platforms-can-learn-reddit's-decentralized-content-moderation

result in a community guidelines violation depending on the situation.[43] Content moderation conducted by users of a platform, as in the Reddit case, addresses the infeasibility of a select team of content moderators accurately reviewing content by distributing the responsibility across platform users. It also combats issues with the misinterpretation of content that stem from artificial intelligence by relying on human beings.

However, placing too much power over content moderation into the hands of users can perpetuate rather than alleviate unfairness in content moderation. For example, my qualitative interview study with Carnegie Mellon University found that over half of the study participants feared that their acceptable content would be reported by platform users and subsequently removed by the platform. One participant explained that users often report content "out of spite" even if it does not violate a platform's community guidelines. Enough spite can turn content moderation into an instance of mob rule, reducing the content moderation process to a tyranny of the majority that casts out unpopular content with sheer force and no justification.[44]

Therefore, the effectiveness of reimagining the technology is *moderate* because although studies suggest that it could address key limitations of human content moderators and algorithms, it risks falling prey to the pitfalls of the reporting process that already perpetuate unfairness in content moderation.

ii) Moral permissibility (high)

One can view social media as a pseudo-nation, with executives who determine laws (in this case, community guidelines) that govern the digital citizens who check "yes" to a platform's terms of agreement. It is common for individual nations to bestow rights to and responsibilities upon their citizens. For example, the United States provides citizens with the right to vote in elections for public officials, while requiring citizens to serve on a jury when they are called upon.[45] This bestowing of rights and responsibilities functions as a social contract, and social contract theorists hold that a person's moral and political obligations depend upon the contract itself, since it forms the society in which one lives.[46] Since social media provides opportunities to its users, including opportunities to connect across vast distances, it appears morally permissible

[43] https://www.forbes.com/sites/kalevleetaru/2019/07/28/the-importance-of-context-and-intent-in-content-moderation/?sh=7ccea67a2a95
[44] Klug, Steen, Yurechko Carnegie Mellon paper (pending publication)
[45] https://www.uscis.gov/citizenship/learn-about-citizenship/should-i-consider-us-citizenship
[46] https://iep.utm.edu/soc-cont/

for platforms to request that users who opt into their services participate in the content moderation process. A social contract would then underly digital rights and obligations, promoting a more just society in the online world.

However, while a social contract is morally permissible, the responsibilities that it places upon members of a (digital) society might not be. In particular, requiring users to moderate content is not an ethical demand. The psychological harm associated with content moderation can be extreme, since the process exposes moderators to depictions of violent and distressing events that can result in secondary trauma that can adversely affect some more than others.[47] This psychological distress hinders the moral permissibility of requiring users to moderate content, since the principle of nonmaleficence holds that there is an ethical obligation not to harm others.[48] It appears ethical, however, to enable users to volunteer as content moderators, just as the United States enables its citizens to engage in jobs like being a police officer that may bring them psychological harm. Granting users the freedom of choices ensures ethicality, so long as social media companies reduce harm to volunteer moderators by providing mental health resources and support.

Therefore, the moral permissibility of reimagining the technology is *high* if it gives users the freedom to choose whether they want to moderate content and provides mental health care if they do engage in the role; otherwise, a requirement to moderate conflict might inflict psychological harm on users, thus violating the moral principle of nonmaleficence.

iii) Political feasibility (moderate)

Though industry lobbying often interferes with the regulation of social media platforms, it is possible that social media companies would support legislation in favor of voluntary content moderation by users. Since content moderation is expensive, with Facebook spending billions to review content each day and platforms including Facebook, YouTube, and Twitter outsourcing content moderation responsibilities to workers at third-party companies, there is a significant

---

[47] https://www.sciencefocus.com/news/content-moderators-pay-a-psychological-toll-to-keep-social-media-clean-we-should-be-helping-them/

[48] https://samples.jbpub.com/9780763773274/Chapter3.pdf?TSPD_101_R0=089de8e4f9ab2000c3c2b85647e401912aad9e893f8ba6862cfcddcc7b311194cd70ca7a73f9b4570815636ef2145000185cdc10b4d0eb9b60f69e0a9090112694cffa1b77eec4f6323966cd79ebd4e9b7faf6f3b3259ed83b558e9b219f395de65d507625c234b95adc81c26d42b8bb14de065c02dfbf48cbd583fcd45816de

financial incentive for companies to rely more strongly on the free, voluntary labor of their users for content moderation.[49]

Therefore, the political feasibility of reimagining the technology is *moderate* because although industry lobbying is a potential obstacle to mandating revisions in content moderation, there are financial incentives for companies to utilize a volunteer-based system.

Fix the System at Large

Fixing the system at large involves combatting opacity in content moderation and increasing platform accountability in regards to content moderation to adjust the contexts in which content moderation operates. Efforts to increase transparency often rely on demanding clearer descriptions of community guidelines, the consequences for not adhering to community guidelines, and the processes that algorithms and human beings follow when moderating content. Such efforts can make information available to users, the general public, or researchers. Increasing platform accountability involves holding platforms liable through fines or similar measures when they wrongly remove appropriate content or fail to remove troublesome content. I will evaluate these examples to judge the practical effectiveness, moral permissibility, and political feasibility of fixing the system at large.

i) Practical effectiveness (high)

Critics of transparency initiatives claim that they distract policymakers from pursuing more effective methods of regulating platforms. However, transparency is a prerequisite for the success of other regulatory measures that seek to combat issues with content moderation. Transparency measures increase knowledge regarding the problems associated with content moderation as well as the most effective techniques for addressing them. Although making more detailed information about content moderation available to users and the general public might motivate companies to improve content moderation, providing researchers with access to platform data and tools is the most promising means of leveraging transparency to disrupt the unjustified silencing of marginalized groups online. With company data, researchers can conduct independent evaluations into the performance of platform content moderation. They can identify areas for improvement that the platforms themselves might have kept secret or simply missed.[14] With transparency, regulatory agencies can then use content moderation data to hold platforms accountable for their inappropriate silencing of marginalized individuals.

[49] https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html

Therefore, the practical effectiveness of fixing the system at large is *high* because transparency could inform policymakers of the best steps to take in regards to improving content moderation while enabling accountability measures, which would hold social media companies responsible for their silencing of marginalized groups through fines or legal ramifications.

ii) Moral permissibility (high)

It is morally permissible to require platforms to increase transparency into content moderation, as long as platforms are not required either to divulge information that unfairly violates the privacy of their users or to surrender intellectual property to which they are legally entitled. More than a morally permissible goal, increased transparency into social media content moderation is an ethical imperative. Without transparency into content moderation, policymakers lack the ability to hold companies accountable for the harm that they inflict on marginalized communities. The promotion of transparency counters a lack of accountability, thus furthering the Aristotelean notion of justice as the correction of what is inequitable.

Therefore, the moral permissibility of fixing the system at large is *high* because it is not only morally permissible but morally required for social media companies to increase transparency into their content moderation and therefore enable greater accountability for the harms that they inflict.

iii) Political feasibility (moderate)

While requirements for social media companies to change their operations (e.g., to removal harmful but legal material from their platforms) are often contested in public discourse and opposed by the companies themselves, transparency measures do not require an immediate shift in company operations but simply request information regarding company practices. This distinction makes transparency legislation significantly more feasible than measures that demand changes in company operations.

However, transparency legislation requires a regulatory structure to interpret and enforce transparency requirements. For example, an agency must ensure that disclosed platform information is complete and accurate, which might involve an analysis by groups outside of the political realm such as academics.[50] Without an authoritative regulatory

[50] https://www.brookings.edu/blog/techtank/2022/05/10/transparency-is-the-best-first-step-towards-better-digital-governance/

structure to analyze and determine adherence to transparency requirements, political efforts to increase transparency into content moderation risk being empty measures.[51]

Therefore, the political feasibility of fixing the system at large is *moderate* because transparency legislation would not encounter the severity of industry pushback that change-mandating policies face, yet implementing a transparency-oriented policy would require the creation and maintenance of a sophisticated regulatory structure.

Reimagine the System at Large

Reimagining the system at large focuses on recognizing the incurable flaws in social media companies' current structures. The primary example of fixing the system at large involves dividing social media platforms into smaller platforms that are theoretically easier to manage and regulate. I will evaluate this example to judge the practical effectiveness, moral permissibility, and political feasibility of reimagining the system at large.

i) Practical effectiveness (low)

It may seem that splitting up platforms would facilitate fairer content moderation, since decreasing the scale at which content must be moderated would remove strain from human and algorithmic systems. However, breaking up platforms like Facebook does nothing to mend content moderation directly. In fact, it creates more platforms to regulate, which could make the harms of content moderation even more difficult to address. For example, smaller split-up platforms could decide to abide by different community guidelines, assign different consequences for community guidelines violations, and utilize different algorithms for content moderation. These variables and more are already difficult to regulate within the relatively small number of social media companies that exist today. Breaking up platforms would create more digital realms to oversee without specifically intervening in the processes of content moderation that they employ. In addition, splitting up platforms does not necessarily limit the amount of content that the platforms must moderate unless a specific limit is placed on content creation per user. This means that the scale at which content moderation must operate will not necessarily decrease following platform divisions. Finally, it could take approximately ten years to break up

[51] https://www.brookings.edu/blog/techtank/2022/11/01/transparency-is-essential-for-effective-social-media-regulation/

a massive platform like Facebook. In that time, social media could change drastically and the problems of content moderation could significantly worsen.[52]

Therefore, the practical effectiveness of reimagining the system at large is *low* because breaking up platforms does nothing to address the harms of content moderation directly and could allow for the worsening of such harms until an alternative solution is reached.[vii]

ii) Moral permissibility (low)

Though the decision to split up social media platforms might be decided upon with benevolence, intentionality is less morally relevant than the effect of an action. For example, I may intend to promote my dog's wellbeing by feeding him delicious chocolate. But by feeding my dog chocolate, I cause him sickness and pain, which threatens the ethicality of my action since I bring my dog harm. Similarly, there is reason to believe that dividing up social media platforms would bring about harm. For example, Facebook has grown to house more users (2.96 billion monthly active users, to be exact) than the population of any nation on the planet.[53] In splitting up a company like Facebook, policymakers must decide how to divide a plethora of users into distinct domains. The initiative could become a forced digital immigration project, severing international friendships and online familial connections or leaving people with haphazard subsets of their prior social groups, depending on how the division occurs. In addition to harming social media users, breaking up platforms could serve as a distraction from the ethical responsibility to rectify the silencing of marginalized groups online. [54]

Therefore, the moral permissibility of reimagining the system at large is *low* because splitting up platforms could sever human connections and divert attention from an ethical obligation to remedy the unfairness in content moderation.

iii) Political feasibility (low)

The potential breakup of social media platforms is greatly opposed by platforms themselves. For example, in 2020, leaked audio of Mark Zuckerberg made clear that he would sue over Elizabeth Warren's plan to break up social media companies if she were to be elected president.[55] In addition, a key historical example demonstrates the infeasibility of breaking up Big Tech. After years of fighting in court, the federal government won a ruling to split up

---

[52] https://hbr.org/2020/09/breaking-up-facebook-wont-fix-social-media
[53] https://s21.q4cdn.com/399680738/files/doc_financials/2022/q4/Meta-12.31.2022-Exhibit-99.1-FINAL.pdf
[54] https://www.wired.co.uk/article/break-up-big-tech-anticompetition
[55] https://www.nytimes.com/2019/10/01/us/politics/elizabeth-warren-mark-zuckerberg-facebook.html

Microsoft in 2000. However, the ruling was ultimately overturned on appeal since Microsoft disliked the potential split. The George W. Bush administration then decided to settle the case without pushing for its revisitation.[56]

Along with platform pushback, policymakers must also navigate the obstacle of existing competition law. Antitrust experts explain that social media companies must have violated laws to warrant breaking them up; politicians cannot enact a split-up just because it might be a good idea.[57] While some claim that social media companies gained their market power by acting in a predatory way, buying up rival companies to solidify their dominance over social media, others hold that politicians need to rewrite existing competition law since it requires proving that consumers are being charged too much. When Facebook and TikTok allow free access to their products, it is seemingly impossible to establish this claim.[5859]

Therefore, the political feasibility of reimagining the system at large is *low* because a split-up of social media companies would face powerful industry pushback and the need to reexamine or rewrite existing competition law.

My Proposed Solution

The table below summarizes my assessments of the practical effectiveness, moral permissibility, and political feasibility of the four categories of solutions that I evaluated above.

	Fix the Technology	***Reimagine the Technology***	***Fix the System at Large***	***Reimagine the System at Large***
Practical Effectiveness	Moderate	Moderate	High	Low
Moral Permissibility	Moderate	High	High	Low
Political Feasibility	Low	Moderate	Moderate	Low

As the table demonstrates, the most promising intervention into the silencing of marginalized voices via content moderation involves fixing the system at large. However, there

[56] https://www.cnbc.com/2019/05/09/facebook-should-not-be-broken-up-commentary.html
[57] https://www.bu.edu/articles/2019/break-up-big-tech/
[58] https://www.nytimes.com/2020/12/09/technology/facebook-antitrust-monopoly.html
[59] https://www.wired.co.uk/article/break-up-big-tech-anticompetition

are two key obstacles to implementing this intervention: 1) the challenge of creating an effective regulatory body to analyze and determine adherence to transparency requirements, and 2) the lobbying of social media companies against governmental regulation. My proposed solution aims to address these challenges while expanding our perception of fixing the system at large to include both fixing the technology and reimagining the technology. I will also explain how fixing the system at large will bring additional benefits related to the other three categories of solutions.

The Solution

My solution entails creating a policy to promote transparency into content moderation algorithms. The policy involves establishing a Platform Transparency Office (PTO) within the Federal Trade Commission, comprised of algorithms and technology experts, policy specialists, equitable technology advocacy groups like the Center for Humane Technology, and others who can provide unique insights into what platform transparency requires as well as help facilitate transparency efforts. The PTO would be built on the model of the global LEED (Leadership in Energy and Environmental Design) certification program. Just as LEED incentivizes companies and individuals to adopt green building standards by awarding LEED ratings, the PTO could incentivize social media companies to practice algorithmic transparency by awarding platforms an Algorithmic Transparency Certification, enabling the PTO to support platforms' business models of sustaining user engagement amongst backlash over algorithmic opacity.

The first step in practicing transparency would be for social media companies to provide the source code for their algorithms to selected researchers in exchange for an Algorithmic Transparency Certification. The PTO could open a request for proposal to research universities, selecting eight to ten institutions that present best concepts and strategies for conducting algorithms research. Selected universities would sign legal documents to protect the source code, which they would receive via privacy-protected pathways, and would publish annual algorithms reports to improve algorithmic transparency as well as identify interventions into issues like algorithmic bias. Dividing the Algorithmic Transparency Certification into designation levels based on the number of researcher interventions that platforms adopt will encourage adherence to researcher interventions.

Ultimately, this policy was designed with resistance in mind by facilitating opportunities for researchers who will encourage social media companies to rectify algorithmic harms, fostering a more just digital realm.

Addressing Challenges to Implementation

As I mentioned above, there are two primary obstacles to implementing solutions under the category of fixing the system at large: 1) the need to create a sophisticated, authoritative regulatory structure to analyze and determine adherence to transparency requirements, and 2) industry lobbying, which has stalled various policies aimed at regulating social media platforms, including the Platform Accountability and Transparency Act that was introduced in 2021 to increase transparency into social media company practices. I will now explain how my policy effectively responds to these two challenges.

1) Regulatory Structure

My policy works around two important obstacles to establishing an effective regulatory body for overseeing platform transparency efforts.

First, it navigates the issue of creating a regulatory structure in the first place by appending another office to the existing Federal Trade Commission. According to the FTC website, the commission's mission involves "protecting the public from deceptive or unfair business practices and from unfair methods of competition through law enforcement, advocacy, research, and education."[60] The FTC has released a report containing information that is directly related to the digital silencing of marginalized communities. The report commented on the perils of using artificial intelligence to combat issues with content moderation, including the potential of artificial intelligence to discriminate against protected classes of people or unjustly over-block content.[61] These interests of the FTC coupled with its existing structure of offices, such as the Office of Technology that seeks to promote technical knowledge across the FTC, underscore the feasibility and effectiveness of establishing the PTO.[62]

Second, the proposed PTO brings together experts from a diversity of backgrounds including technology, policy, and advocacy to help the office stay up-to-date on the ever-

[60] https://www.ftc.gov/about-ftc/mission
[61] https://www.ftc.gov/news-events/news/press-releases/2022/06/ftc-report-warns-about-using-artificial-intelligence-combat-online-problems
[62] https://www.ftc.gov/about-ftc/bureaus-offices/office-technology

changing social media and policy landscapes. Like the existing Office of Technology, which brings together subject matter experts across artificial intelligence, human-computer interaction, and social science as it relates to technology, the PTO brings a variety of skilled thinkers to the table to address the complex, interdisciplinary issue of unfair silencing on social media.[31] Historically, the science and policy realms have been relatively separate. Authors Marta Sienkiewicz and David Mair of the Science for Policy Handbook describe the Science-Policy Binary Separation, arguing for a movement away from Science for Policy 1.0 through which detached scientific advice is given to policymakers ad hoc to Science for Policy 2.0, which involves a united, collaborative system of scientists, policymakers, and other stakeholders.[63] Such collaboration across science and policy would enable the PTO to review bids from institutions that wish to receive access to social media algorithms, while ensuring that the selected institutions receive the algorithms in privacy-protected ways. It would also support the PTO in determining whether social media companies comply with the criteria needed to receive various levels of the Algorithmic Transparency Certification.

2) Industry Lobbying

My policy circumnavigates the predominant obstacle to regulating social media platforms: the resistance of the platforms themselves to regulation. In 2022, the Digital Services Act became Europe's first major legislation requiring online services to provide data to researchers. Similar acts in the United States, including the Platform Accountability and Transparency Act, have failed to address the problem—they *compel* social media companies to make their algorithms more transparent rather than *incentivize* them.

The Platform Accountability and Transparency Act is described as "requiring social media companies to share data with the public and researchers so that we can look under the hood and finally see a clear picture about the effects these platforms have on all of our lives."[64] Though this aim is noble, the lobbying powers of Big Tech are only increasing resistance to such a goal. For example, the tech industry group NetChoice spent $170,000 on lobbying in just the

[63] https://www.sciencedirect.com/science/article/pii/B9780128225967000012
[64] https://www.coons.senate.gov/news/press-releases/senator-coons-colleagues-introduce-legislation-to-provide-public-with-transparency-of-social-media-platforms

first quarter of 2022 to push back on legislation aimed at holding social media companies accountable for inaccurate content moderation.[65]

In the face of industry power and failing efforts at demanding platform compliance, my proposed method of incentivizing rather than requiring a shift in platform behavior is promising. However, I do acknowledge that additional efforts might be necessary to encourage social media companies to opt into providing their algorithmic source code to research institutions in exchange for an Algorithmic Transparency Certification. My outlined approach relies on hard power, using reward to incentivize, but might require soft power as well. For instance, it may be necessary to elucidate the potential benefits of the aforementioned exchange (e.g., the satisfaction of increasingly conscientious and algorithmically literate users) to social media companies.[12]

Additional Benefits

By forging partnerships between computing researchers and social media companies, my policy holds the power to increase the effectiveness of two other categories of solutions that I outlined. In particular, it can push the boundaries of fixing the technology and reimagining the technology. The provision of algorithmic source code to researchers will enable them to delve into the black boxes that were previously unavailable to them. For example, research teams can develop particular test sets to feed into a given platform's content moderation algorithm, enabling them to see how exactly the algorithm responds to their test data. In practice, this might involve researchers creating a set of posts from Black users, examining which posts are wrongly flagged by the algorithm as violating the community guidelines, and pinpointing reasons for the demonstrated bias that are buried in the model. In summary, my policy does not only make issues with content moderation more transparent and therefore approachable by skilled researchers; it specifically puts researchers on a mission to remedy these issues at no cost to social media companies themselves, thus providing the companies with free expertise into how they can appease an increasingly conscientious user base with an agitated population of marginalized users. My proposal also breaks through the widespread barrier to regulating social media platforms through an incentivization-based approach, potentially paving the way for a long-lasting and mutually beneficial relationship between researchers and social media companies that

[65] https://www.opensecrets.org/news/2022/06/big-tech-groups-ramp-up-lobbying-amid-fight-over-social-media-content-moderation/

can enable other regulatory measures. Finally, my proposal touches on the fourth category of solutions that I described above (i.e., reimagining the system at large). Without propagating the harms associated with splitting up platforms, it achieves the goal of reimagining the architecture of the social media world by giving smaller social media companies a chance to compete with larger platforms. This is because smaller companies can make their algorithms transparent to researchers and obtain an Algorithmic Transparency Certification that is visible and publicized. Much like how approximately eighty million people elect to use DuckDuckGo instead of Google, Bing, or other predominant search engines because of its prioritization of user privacy, users will likely appreciate the transparency emphasis of smaller social media companies that opt into my policy, thus making those companies legitimate competitors with larger social media platforms.[66]

Conclusion

The disproportionate silencing of marginalized groups online is a difficult issue to comprehend, given the technical language of content moderation systems and opacity in which they are shrouded. Throughout my paper, I elucidate the policies and enforcement mechanisms that underly social media content moderation, then discuss what causes content moderation to impact marginalized communities negatively and unduly. I describe four categories of solutions to issues of bias against marginalized groups on social media: *fix the technology*, *fix the system at large*, *reimagine the technology*, and *reimagine the system at large*. Then, I evaluate the *practical effectiveness*, *moral permissibility*, and *political feasibility* of these four categories of solutions, arriving at my own solution to evaluate its efficacy, ethicality, and feasibility in comparison to the four categories of solutions that I analyzed.

My proposed solution provides a glimmer of hope amidst an ongoing, largely stagnant, and defeating battle against the injustices of Big Tech. However, it is simply the first step in addressing the issues posed by these companies, which, given the scope of this paper, I illustrate through the lens of the disproportionate silencing of marginalized groups online. Further steps must be taken to continue to protect the dignity and fundamental equality of marginalized individuals in digital spaces. My policy seeks to pave the way for these future efforts by increasing transparency into social media algorithms. It does this not only for researchers so that they can investigate the source code and pose interventions, but it makes traditionally covert

[66] https://spreadprivacy.com/how-many-people-use-duckduckgo/

social media operations more visible and understandable to the public, including marginalized users themselves. By requiring the publication of annual reports from researchers that give greater access to the often elite and exclusionary language of the tech world, my policy attempts to level a historically uneven playing field for people who want to engage in conversations around technological injustice. Through an increase in users' agency brought about by a potential for algorithmic revisions and a guarantee of greater transparency, opting into my policy would begin to transform social media companies into truly democratic spaces in which users' opinions and lived experiences can be shared without discrimination, given that they do not bring harm to others.

---

[i] The Partnership for Countering Influence Operations (PICO) exists within the Carnegie Endowment for International Peace, which was developed in 1910 with a $10 million gift from Andrew Carnegie. It is a nonpartisan international affairs think tank comprised of more than 300 people across twenty countries who offer insights to policymakers to facilitate global peace. The AllSides political bias assessment company gives PICO a bias rating of "center," meaning that it does not show much media bias or it balances left and right biases.

[ii] Twitter, Facebook, and YouTube tend to have Particularized Policies regarding unacceptable behavior, whereas TikTok, Pinterest, Reddit, Tumblr, LinkedIn, Instagram, Gab, Telegram, WhatsApp, and Signal tend to have Generalized Policies.

[iii] Instagram has not been formally pressed on the brevity of its community guidelines, nor have writers publicly and reliably commented on the vagueness of its rules. However, philosophers have commented on the nuances of hate speech among other concepts that platforms usually seek to moderate, which suggests that the concepts themselves are philosophical in nature and require often complex definitions.

[iv] The realization that social media algorithms can disproportionately and unfairly silence marginalized users is a recent discovery in the research world, dating back to the 2010s. Due to the recency of this realization and the small number of papers that investigate the algorithmic silencing of marginalized groups, it is difficult if not impossible to gauge whether social media algorithms have become less discriminatory over time.

[v] Few bodies have imposed fines on social media companies for reasons related to content moderation. A recent example is a new Florida law, which fines social media companies that permanently ban political candidates (https://www.nytimes.com/2021/05/24/technology/florida-twitter-facebook-ban-politicians.html). Other measures to hold social media companies accountable for their harms (e.g., paying the victims of their harms) have been initiated through lawsuits. However, they have not yet been addressed at a deeper level that attempts to confront the causes of the harms at their source.

[vi] Specifically, my conversations with Dr. Levy, Dr. Goldberg, and Dr. Pickett contribute to my analyses of the four categories of solutions that I describe, which I then utilize to evaluate my proposed solution.

[vii] My low ranking of the practical effectiveness of reimagining the system at large could be more nuanced or reversed if alternative arguments are employed. However, from my synthesis of trustworthy journalistic sources, I deem the effort to be of low practical effectiveness.