Lip(s) Service: A Socioethical Overview of Social Media Platforms' Censorship Policies Regarding Consensual Sexual Content

Sage Futrell
William & Mary

Lip(s) Service: A Socioethical Overview of Social Media Platforms' Censorship Policies Regarding Consensual Sexual Content

Sage Futrell (W&M '26)

Coastal Virginia Commonwealth Cyber Initiative

Cybersecurity Undergraduate Research Program

Dr. Michael Lapke

Table of Contents

I. Sealed Lips: A Legal Introduction to Sexual Censorship in Cyberspace

Social media platforms are undeniably some of the fastest growing entities on the Internet. Over 4.6 billion people are active on social media, with continued growth even decades after their initial rollout (Hootsuite, 2022).[1] With this rapid rise in user base comes a variety of challenges related to moderating the content they post online. Regulating and combatting online sexual exploitation and trafficking is an issue that has received widespread attention from the media and the government. In 2016, human trafficking hotline reports increased by nearly 35%[2], and it is reported that as many as 70% of child sex trafficking victims were recruited via social media platforms.[3]

Fight Online Sex Trafficking Act/Stop Enabling Sex Traffickers Act (FOSTA-SESTA) was a law passed in 2018 dedicated to addressing this issue.[4] Specifically, FOSTA-SESTA was a response to backlash surrounding Backpage.com's controversies involving illegal sex work advertisements. In 2017, a Senate investigation revealed that Backpage had been "complicit in obscuring ads for child trafficking."[5] Shortly after, a well-known documentary titled *I Am Jane Doe*, which detailed the website's infamous involvement in trafficking, was released. This caught the attention of Congress, and FOSTA-SESTA was developed.

The law serves as an updated version of the 1996 Communications Decency Act (CDA), which set general guidelines for user safety and privacy on Internet Service Provider platforms. Specifically, Section 230 of the CDA was deemed "outdated" by the Trump administration and was the reason why Backpage was previously dismissed of its charges. According to Section 2 of FOSTA-SESTA, "[1996 Section 230] was never intended to provide legal protection to websites that unlawfully promote and facilitate prostitution and websites that facilitate traffickers in advertising the sale of unlawful sex acts

[1] https://www.hootsuite.com/resources/digital-trends

[2] https://www.nbcnews.com/news/us-news/human-trafficking-increased-2016-organization-reports-n717026

[3] https://www.huffpost.com/entry/sex-trafficking-in-the-us_n_5621481

[4] https://www.sciencespo.fr/public/chaire-numerique/en/2023/06/30/student-essay-how-us-law-fosta-sesta-inadvertently-censored-sex-workers-social-media-usage/

[5] https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom

with sex trafficking victims; (2) [such sites] have been reckless in allowing the sale of sex trafficking victims and have done nothing to prevent the trafficking of children and victims of force, fraud, and coercion; and (3) clarification of such section is warranted to ensure that such section does not provide such protection to such websites."[6]

FOSTA-SESTA was meant to hold ISPs accountable for user's sexual content – but it didn't specify how *consensual* content related to sexuality should be handled. By misconstruing all sex work and expressions of sexuality with illegal activities, this law set the stage for major social media platforms to enact biased community guidelines that target sex workers, LGBTQ+ creators, and other marginalized groups.[7] Although the Internet has a long and convoluted history of censorship, many disillusioned users cite this one specific law as the catalyst for their troubles related to maintaining online engagement.

There has been an extensive amount of previous literature published that has examined the impact (or rather, lack thereof[8] [9] [10]) of FOSTA-SESTA on online sex trafficking and similar crimes. Rather than focus on the narrative of illegal sexual activities, this paper aims to look at how this law shaped consensual sexuality on social media. As a content moderator myself, I will explore a smaller platform that I conducted action research on, which was created as a response to stricter censorship guidelines regarding sexuality: Lips.

Founded by Annie Brown (W&M '11), Lips.social is a blockchain-enabled social media website that relies solely on human content moderators; however, an AI content moderation system, which could filter creator applications and efficiently manage the site's homepage, is in the works. The use of deep-

[6] https://www.congress.gov/bill/115th-congress/house-bill/1865/text
[7] https://www.aclu.org/news/civil-liberties/how-online-censorship-harms-sex-workers-and-lgbtq-communities
[8] https://whyy.org/segments/fosta-sesta-was-supposed-to-thwart-sex-trafficking-instead-its-sparked-a-movement/
[9] https://www.theverge.com/2021/6/24/22546984/fosta-sesta-section-230-carveout-gao-report-prosecutions
[10] https://www.fastcompany.com/90741323/four-years-after-sesta-fosta-a-new-bill-investigates-its-harm

learning AI-powered social media algorithms is not new, and such systems are often considered culprits of over-censoring sexual expression.[11]

As a content moderator intern for Lips, I asked myself: to what extent should I censor content on this explicitly anti-censorship platform? What biases do existing AI algorithms on other social media hold against certain marginalized groups? Can community-driven data labeling help AI do the same work that I and other human mods can, such as assessing the artistic merit of erotic content?

This paper aims to answer these questions using case studies of social media platforms and their histories of online censorship, specifically that of content related to sexuality.

Conceptual Framework & Definitions

This paper is primarily concerned with the impact of changing content moderation policies on the safety and security of sexual creators. Many of the opinions expressed will be pro-SW/sex positive. Much has already been said on the state of human trafficking/sexual abuse online, so I aim to fill the research gap regarding consensual sexuality on social media.

A legal and sociological-ethical ("socioethical") framework is employed when analyzing qualitative data from both literature reviews and action research in this paper. Social factors are considered with an intersectional and holistic approach – race, gender, and sexuality are perceived together in a variety of ways online, just as they are in real life. In Cornell University's computer science publication *arXiv*, researcher Jamell Dacon states that it is necessary to acknowledge the importance of "holistically addressing pressing concerns of AI systems from a socioethical impact assessment perspective to explicate its harmful societal effects to truly enable humanity-centered Trustworthy AI."[12]

[11] https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies
[12] https://arxiv.org/abs/2309.09450

Trust and Safety (T&S), which entails practices and tools that promote online security and fairness,[13] acts as the ethical foundation of my action research as a content moderator.

This paper will discuss deep machine learning in artificially intelligent (AI) algorithms on social media, and how T&S practices can be incorporated into AI/Natural Language Processing (NLP) and Large Language Models (LLM) to reduce misflagged content. Some AI algorithms are noted to use computer vision, which is described in literature as "a deep learning process which uses neural networks to decipher images within a photo or video."[14]

Specifically, the use of human-in-the-loop (HITL) systems as a potential solution to biased social media algorithms will be explored. To understand HITL, I must first explain its opposite, the Big Red Button (BGG) system. The Stanford Institute for Human-Centered Artificial Intelligence describes BGG as "a technology that reliably delivers the right answers while hiding the process that leads to them." In other words, an entirely automated AI system that takes no adaptive input from human sources and instead functions based on predetermined settings. For example, an AI art generator may be programmed to copy a certain artist's style, but it doesn't understand the emotional connotations and meaning behind that style. HAI proposes the concept of "selective human participation" in AI development. By keeping people "in the loop," content can be curated in more intentional ways and biases can be continuously checked.[15]

Shadowban is a recurring term with many meanings; but within this paper, it holds a specific context. In the early days of the Internet, this act went by many names – "stealth banning," "ghosting," and "hellbanning" – and was a method employed by human content moderators on small forums and chatrooms as a way to censor certain users without those users knowing.[16] In contemporary context,

[13] https://www.spectrumlabsai.com/trust-and-safety#:~:text=Contact%20Us%20Today-,What%20is%20Trust%20and%20Safety%3F,that%20are%20outside%20community%20guidelines.
[14] https://dev.to/mage_ai/how-does-tiktok-use-machine-learning-5b7i
[15] https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems
[16] https://www.vice.com/en/article/a3q744/where-did-shadow-banning-come-from-trump-republicans-shadowbanned

shadowbanning is done by algorithms – certain keywords on a user's post may automatically keep it off of the reels, feed, or explore pages of non-followers. This paper will focus on the phenomenology of TikTok and Instagram's shadowbans specifically, based on my action research work on Lips' official IG page and the experiences of other users.

Methods

A variety of media (articles, academic studies, podcasts, etc.) were reviewed and cited in the "Background" and "Case Studies" sections of this paper. Much of the literature discussed in this paper was created between the years 2017 (shortly before the passing of FOSTA-SESTA) and the current year, 2023. For each social media case study, the following questions were considered:

- Demographics: What kinds of people make up the user population on this platform, and what types of content are shared amongst them? How has the hegemony of this platform changed/stayed the same over time?

(Note: Case studies only cover social media sites from a Western, English-speaking lens, given the fact that FOSTA-SESTA was passed in the US and much of the literature in this paper references its impact on American users.)

- Moderation methods: Does this platform use automated/AI or human mods or a mixture of both? If applicable: How many human mods regulate content? To what extent are automated/AI mods trained to judge content contextually?
- Censorship history: How did FOSTA-SESTA change this platform's community guidelines and content moderation methods? What are some instances of over/under-censoring of certain content related to sexuality?
- User interactions: How have users responded to changes in guidelines? How much communication exists between users, moderators, and developers?

In the second part of each case study, I included a description of the action research I performed on each platform. For example, Lips has an official Instagram account that I was tasked with managing. Although Lips is also active on META's Threads, I have decided to exclude this platform, given how new it is.

My action research was performed during my internship with Lips, which spanned one semester (14 weeks). My general tasks as an intern included content moderation, data organization, and social media management. I also observed developing team meetings regarding Reliabl and helped with organizing data for outreach workshops. Most of the action research data presented in this paper is sourced from Lips.social's administration page for creator applications, Lips' official Instagram page, Tumblr page, and TikTok account.

II. Lip[s] Service: Current Literature & Action Research Case Studies

Case Study: META (Instagram & Facebook)

Based on numbers of users, META owns the largest contemporary social media platforms. According to a 2022 survey, "77% of Internet users, about 3.59 billion people, are active on at least one META platform."[17] An estimated 15k content moderators have been employed to monitor content on both Instagram and Facebook. For reference, that is around 1 content mod for every 333k users.[18]

The *MIT Technology Review* notes that the majority of META's 15k human content moderators are contract workers employed by third-party entities. Facebook representatives have admitted that these mods have an approximately 10% error rate on their site, and many fail to recognize community guidelines violations on account of personal biases.[19] New York University's Stern Center for Business and Human Rights reports that to lower this error rate, META would have to employ twice as many human mods from diverse cultural backgrounds and give them full-time benefits. As author and deputy director of the center Paul M. Barrett puts it, "Content moderation is not like other outsourced functions, like cooking or cleaning. It is a central function of the business of social media, and that makes it somewhat strange that it's treated as if it's peripheral or someone else's problem."[20]

Rather than hire more human workers and centralize its moderation methods, META has made advancements to AI algorithms that can take the jobs of its human workforce and cover more ground. Since its establishment, Instagram has been a trendsetter for deep learning algorithmic systems. After being acquired by META, Facebook's text analytics program DeepText was incorporated into Instagram's feed, helping filter spam and present targeting advertisements to users. Shortly before the

[17] https://www.investing.com/academy/statistics/facebook-meta-facts/#:~:text=More%20than%2077%25%20of%20Internet,by%202%2C203%25%20over%20ten%20years.
[18] https://www.nytimes.com/2018/12/27/world/facebook-moderators.html
[19] https://www.technologyreview.com/2020/06/08/1002894/facebook-needs-30000-of-its-own-content-moderators-says-a-new-report/
[20] https://bhr.stern.nyu.edu/tech-content-moderation-june-2020

passing of FOSTA-SESTA, Meta used AI to perform various studies "on the human condition."[21] One project involved data-mining 100 million photos of global clothing textiles as a way to introduce the platform's algorithm to cultural analysis.[22] Although Meta has put effort into making its algorithm understand human differences, a number of issues arose when it was tasked with detecting sexual content.

After FOSTA-SESTA, Instagram updated its Terms & Guidelines[23] with the following: "we don't allow nudity on Instagram… photos, videos, and some digitally created content that show sexual intercourse, genitals, and close-ups of fully-nude buttocks…offering sexual services [is] also not allowed." When asked by users to elaborate, META releases a statement saying that they "allow for the discussion of sex worker rights advocacy and sex work regulation" but "draw the line… when content facilitates, encourages, or coordinates sexual encounters or commercial sexual services between adults." The platform's representatives state that doing this will prevent potential trafficking and coercion.[24]

META CEO Mark Zuckerberg agrees on the importance of recognizing "nuance" in content moderation, but he believes that his company's AI is capable of context-based moderation. In a hearing on Facebook's problems with misinformation, Zuckerberg revealed that a staggering 95% of flagged content is detected by AI-powered algorithmic moderators.[25]

Although META "commend[s] the [expression of] desire for sexual activity, promoting sex education, discussing sexual practices or experiences, or offering… programs that teach techniques or discuss sex," there is substantial evidence of this claim being hypocritical.

---

[21] https://www.forbes.com/sites/bernardmarr/2018/03/16/the-amazing-ways-instagram-uses-big-data-and-artificial-intelligence/?sh=1037bac95ca6
[22] https://www.technologyreview.com/2017/06/15/105762/data-mining-100-million-instagram-photos-reveals-global-clothing-patterns/
[23] https://help.instagram.com/477434105621119
[24] https://transparency.fb.com/policies/community-standards/sexual-solicitation/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fsexual_solicitation
[25] https://www.businessinsider.com/zuckerberg-nuances-content-moderation-ai-misinformation-hearing-2021-3

For example, the Center for Intimacy Justice studied 60 online businesses focused on women's sexual and reproductive health, and all said that at least one of their Meta ads had been rejected and half had their whole accounts suspended. Rather than analyze each SRH account's advertisements individually, Facebook's algorithm was found to automatically take down multiple ads at a time, including ones that did not violate any guidelines.[26]

There have been times when Instagram edited its policies after facing significant backlash from users for their censorship rules. When model Nyome Nicholas-Williams had some of her photos taken down within hours of being posted online, she started a campaign that pointed out how Instagram's policies on non-sexual nudity can disproportionately affect plus-sized individuals. Her followers rallied behind the hashtag #IwanttoseeNyome and quickly gained mass attention. On top of fatphobia, many also viewed this as a case of anti-Black bias, as users noted that white models who posted similar risque photos didn't face the same post deletions as Nicholas-Williams and others had. Adam Mosseri, the CEO of IG, gave a response saying that he "[heard] concerns about whether we suppress Black voices and whether our products and policies treat everyone equally." Shortly afterwards, he and vice-president Vishal Shah launched a campaign to address internal biases on the site, #ShareBlackStories. They also reinstated Nicholas-Williams's posts and issued a formal apology to her.[27]

However, not all voices on this issue have been heard. Smaller creators find themselves facing the worst of post-FOSTA-SESTA censorship policies. Mistress Marley, whose "main source of income" was Instagram, had her accounts deleted several times within the span of three months. She claims that when she used certain hashtags that alluded to her profession, her account would be shadowbanned and she couldn't maintain followers. After repeatedly trying to appeal Instagram's decisions, Marley eventually gave up on using social media to advertise. The video series she speaks on, "SEX FACE," highlights how

[26] https://www.ohchr.org/sites/default/files/documents/issues/digitalage/cfis/tech-standards/subm-standard-setting-digital-space-new-technologies-csos-choice-rnw-media-3-input-part-2.pdf
[27] https://www.theguardian.com/technology/2020/aug/09/instagrams-censorship-of-black-models-photo-shoot-reignites-claims-of-race-bias-nyome-nicholas-williams

"many sex workers… experienc[e] an increase in anxiety when engaging online for fear of having their platforms deleted, flagged, or shadowbanned."[28]

A similar story comes from Natassia Dreams, who has created 12 new Instagram accounts after her first one was banned after 10 years.[29] Her first account had amassed 70k followers over 10 years of activity, but now, Dreams struggles to get past 700 followers without her account getting suspended.

Dreams' testimonial is very telling of the "shadowban phenomenon" on Instagram. Based on the sheer amount of attention Instagram's censorship has gotten in the media, the very term "shadowban" has become associated with the platform. The most recent definition is that a shadowban doesn't remove posts or delete users, but instead discreetly "conceals [content] from public view," preventing users from expanding their reach.

Mark Zuckerberg instead refers to shadowbanning as the identification of "non-recommendable" content, which he doesn't attribute to algorithmic biases but rather "mistakes" made by improper flagging. [30]Adam Mosseri, in contrast, has offered a more ambiguous interpretation of the platform's censorship methods. In a 2023 blog post,[31] he stated: "Contrary to what you might have heard, it's in our interest as a business to ensure that creators are able to reach their audiences and get discovered so they can continue to grow and thrive on Instagram."

When Lips.social has started the online campaign #bannedonthegram, over 250k IG accounts used the hashtag to talk about their experiences with censorship on the platform. My own experiences managing the @lips_zine IG account allude to symptoms associated with the platform's shadowban.

I began my internship with Lips just as the platform's founder was reconsidering Instagram as a means of advertising. Despite regularly posting and engaging with others, Lips' Instagram account had

[28] https://youtu.be/ZO1CsWRXEcE?si=J9YLv94YnOcKKOTm
[29] https://schoolofsexed.org/blog-articles/2020/09/17
[30] https://www.businessinsider.com/mark-zuckerberg-no-shadow-ban-facebook-but-mistakes-are-made-2022-8
[31] https://www.instagram.com/reel/Cs6gh_NgPF0/?utm_source=ig_embed&ig_rid=daac66e5-112d-4b18-bd78-0e271570bcd2

seen a rapid drop in user engagement and a steady decrease in followers over the past few months. When looking at Insights data from the past 90 days, the numbers are jarring:

1. 

During my internship period, our following dropped from 23.3k to 23.1k.

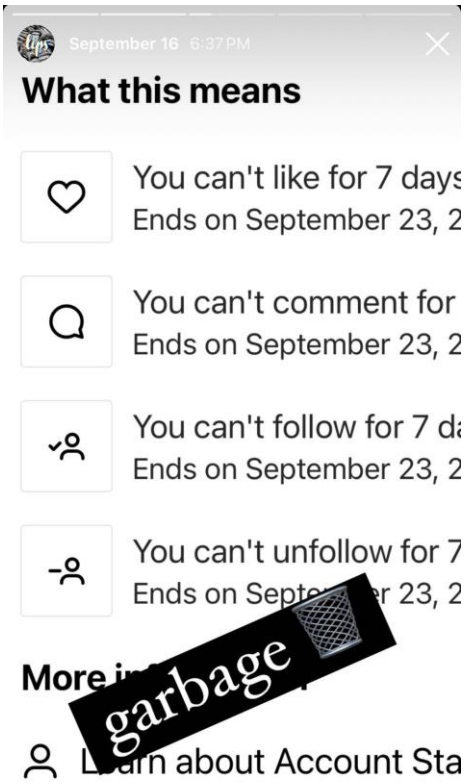
I've also taken note of our lack of reach in regards to Stories. On average, only about 115 different accounts – most being from dedicated followers – see Stories from @lips_zine. This is a stark contrast to the account's 23k followers and suggests some classic "shadowbanning."

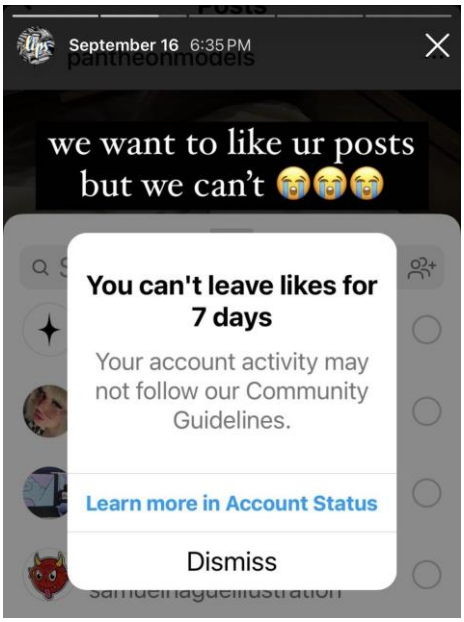
On top of this, the @lips_zine account has been under constant threat of permanent suspension. Lips's page primarily serves as a place to spotlight marginalized creators on IG, particularly queer artists and photographers. However, due to the nudity and erotic themes present in the artwork Lips reposts, the account is prone to being misflagged for sexual solicitation.

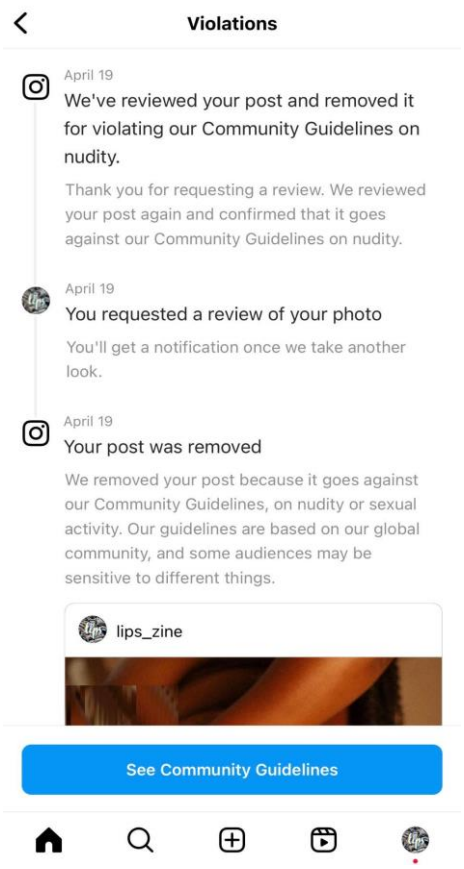
As shown in the images below, there was a period of time during my internship when @lips_zine could not like, comment, or follow other content for seven days. Presumably, this was caused by a creator highlight that was posted on September 6, which features art from @xenotrip that contained animated nudity.

This was something I had never personally seen happen to an account before, and it was also news to my internship director. Previously, the official Lips account had only been periodically notified of posts violating Instagram's guidelines on nudity/sexual content (see image 4). All I could do was post on the account's story, protesting Instagram's decisions.

Later on, I posted less on Instagram to avoid further suspensions. As of writing this report, the account is still up but Lips has shifted its attention to marketing on TikTok (which will be discussed later).

2. 

3. 
September 16 6:35 PM
pantheonmodels
we want to like ur posts but we can't 😭😭😭
You can't leave likes for 7 days
Your account activity may not follow our Community Guidelines.
Learn more in Account Status
Dismiss

4. 
Violations
April 19
We've reviewed your post and removed it for violating our Community Guidelines on nudity.
Thank you for requesting a review. We reviewed your post again and confirmed that it goes against our Community Guidelines on nudity.
April 19
You requested a review of your photo
You'll get a notification once we take another look.
April 19
Your post was removed
We removed your post because it goes against our Community Guidelines, on nudity or sexual activity. Our guidelines are based on our global community, and some audiences may be sensitive to different things.
lips_zine
See Community Guidelines

Case Study: Tumblr

What sets Tumblr apart from other sites is the social hegemony of its platform. Unlike META's platforms, Tumblr has historically been considered a safe space for the queer community, neurodivergent members of fandom, SWers, NSFW artists, and other transgressive minority groups.[32] The site's About page says it all: "It's wholesome chaos. It's the gay people in your phone… Tumblr is whatever you want it to be."[33]

This difference in user-admin interactions expands to how staff describe the platform's moderation landscape: "We moderate content using a mix of machine-learning classification and human moderation from our team of trained experts. Computers are better than humans at the scaling process—and we need them for that—but they're not as good at making nuanced contextual decisions. That's why when you appeal a post we've marked as mature, it gets sent to a real, live human who will look it over with their real, live human eye(s)."[34]

The Staff blog has been used to respond to user concerns, particularly those related to moderation. For instance, when Tumblr implemented Safe Mode in 2017 to filter nudity using a photoset-analyzing algorithm, some users in the LGBTQ+ found that their innocent posts were being incorrectly tagged as mature content.[35] Within a week of adding the Safe Mode feature, staff responded with a transparent overview of how mistagging had become commonplace: "Because we consider Explicit blogs to be *predominantly* sensitive content, we were automatically marking all their posts as sensitive. That was too broad. [Additionally, i]f an *Explicit* Tumblr reblogged a *safe* post, we were marking that reblog as sensitive. This was even happening to text posts. Which is silly."[36] In this same post, they detailed plans to individually analyze posts and revise the algorithm's photoset classification system.

---

[32] https://www.tandfonline.com/doi/full/10.1080/14680777.2019.1678505
[33] https://about.tumblr.com/
[34] https://help.tumblr.com/hc/en-us/articles/360011799473-Content-moderation-on-Tumblr
[35] https://techcrunch.com/2017/06/24/tumblr-says-it-fixed-the-safe-mode-glitch-that-hid-innocent-posts-including-lgbtq-content/
[36] https://staff.tumblr.com/post/162178688374/safe-mode-update

After these fixes, Not Safe For Work (NSFW) communities continued to thrive on Tumblr. SimilarWeb, a site analytics service, "estimates that adult content drives 20.53 percent of clicks to Tumblr's desktop site." For comparison, the second most popular category of content, literature, only brings in 7.61% of clicks.[37] That is, until FOSTA-SESTA.

After this law was passed, Tumblr enacted new community guidelines that are infamously referred to as the platform's "porn ban." This shift coincided with Yahoo's acquisition of the platform and the recent discovery of child exploitation material on the site. The app had been temporarily removed from the Apple app store during this time as developers investigated the appearance of child sex abuse content on Tumblr.[38] Yahoo's CEO aimed to "cleanse" this platform in order to market it to larger audiences and prevent future incidents of illegal content on the site. Once again, what started as a well-intentioned attempt at stopping trafficking ended up censoring a variety of consensual content.

Ironically, this caused a massive loss in user engagement, as nearly one third of Tumblr's user base left the site.[39] Many moved to Twitter, TikTok, or other platforms that would allow their content.

Those who stayed active during this "porn ban" noted almost comical instances of misflagged posts, ranging from fully clothed selfie photos of women to a picture of a vase. As another user summarized, "the new content-ID/flagging system is not quite working properly yet, to put it kindly. In the shadowbans/mini purges, it wasn't child porn or pornbots that were being targeted—it was educational things, SFW things, and completely random things."[40]

Although Tumblr's staff did acknowledge the difficulties that the platform's AI had in correctly identifying content, little was done to stop posts from being wrongly taken down.[41] Some post deletions were even perceived as malicious. For instance, in a viral Tweet, a transmasculine user describes how his

[37] https://techcrunch.com/2017/06/20/tumblr-rolls-out-new-content-filtering-tools-with-launch-of-safe-mode/
[38] https://www.usatoday.com/story/tech/news/2018/11/20/tumblr-ios-app-store-apple-child-pornography/2073740002/
[39] https://www.nytimes.com/2022/11/02/technology/tumblr-nudity-explicit-posts.html
[40] https://kotaku.com/tumblr-porn-ban-leaves-artists-and-fans-seeking-new-pla-1831056412
[41] https://www.cnn.com/2019/01/02/tech/ai-porn-moderation/index.html

post-top surgery photos were removed based on showing "female-presenting" nipples. "So you took down my photos and actively misgendered me," the post reads.[42] Cases like this mirror the sense of alienation that LGBTQ+ users experienced after FOSTA-SESTA brought upon stricter community guidelines.

Even after Tumblr reversed its adult content ban in 2022, many popular communities on the platform were still fragmented. The platform released a "community labeling" system to lessen algorithmic biases, but users quickly found that this system now relied too heavily on user feedback. Having one user tag a post was all it took for it to be categorized by the algorithm – no moderator, human or AI, seemed to look over these tags. For instance, bot accounts got away with posting overt sexual content by labeling them "SFW," while users could take down each other's posts by tagging them as "NSFW."[43] It seems that developers have fixed this flaw, but its initial implementation didn't help the general distrust that users felt towards the platform.[44]

Part of my social media management work for Lips involved reviving their official Tumblr account, @lipszine, which had previously been abandoned shortly after the 2018 adult content ban. During my internship, Tumblr was more of a tool than an active platform that Lips advertised on. I used this site to find creators who may be interested in joining Lips or having an artist feature on our IG page, and from there I reached out to them via Instagram messages. What made this task difficult was the loss of several users who had once been closely associated with Lips' blog. Out of the accounts that the @lipszine Tumblr page follows, I discovered that nearly 40% had gone inactive between 2018-2021. With that said, I did little content creation on this platform and instead focused my energy on Instagram and TikTok.

---

42 https://twitter.com/gachabasta/status/1389758728056950789?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1389758728056950789%7Ctwgr%5Eb22a6965487925b4ccf96d5b03a47c3736d34d34%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fwww.vice.com%2Fen%2Farticle%2F93yyp8%2Ftumblr-is-trying-to-win-back-the-queer-audience-it-drove-off

43 https://www.themarysue.com/is-nsfw-back-on-tumblr-community-labels-on-tumblr-explained/

44 https://www.tumblr.com/angelltheninth/707593013502836736/community-labels-are-killing-fandoms-and-creators?source=share

Case Study: TikTok

In August of 2018, a few months after FOSTA-SESTA was enacted, Chinese company ByteDance acquired the social media site Music.ly and rebranded it into the music-savvy app TikTok.[45] However, it wasn't until the beginning of the COVID-19 pandemic that TikTok would find popularity amongst American audiences for its bite-sized content and highly-specific recommendation page.[46]

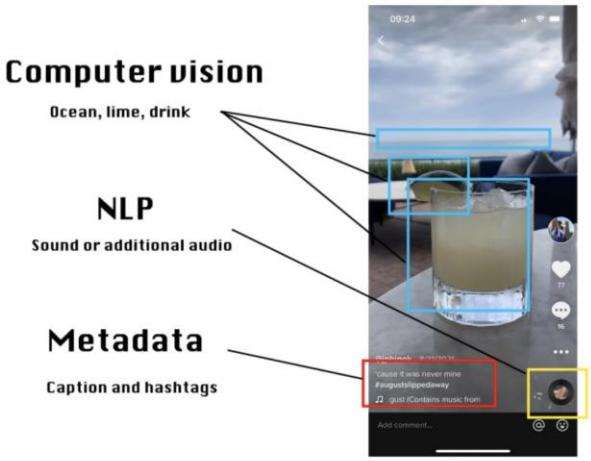
Many users, such as pole dancer and professor Dr. Carolina Are, moved to TikTok in response to stricter community guidelines on other platforms. "I got a TikTok in early 2020 as an experiment," Dr. Are explained in an interview with *Input*. "I wanted to see if it was easier to grow on [the platform] as a pole instructor than it was on Instagram, which has been plagued by shadowbanning and censorship."[47]

Like Instagram's explore page, TikTok's For You Page (FYP) and its use of deep learning technology was initially popular amongst users. The platform's addictive algorithm, which pushes short, palatable videos to like-minded audiences, made it the most popular download of 2021.[48]

According to TikTok, the FYP is "a stream of videos curated to your interests, making it easy to find content and creators you love… powered by a recommendation system that delivers content to each user that is likely to be of interest to that particular user."[49]

TikTok's AI algorithm examines three factors when recommending content to users: "computer vision, [NLP], and metadata."[50] The following figure depicts how each method is employed when analyzing video content:

[45] https://www.theverge.com/2018/8/2/17644260/musically-rebrand-tiktok-bytedance-douyin
[46] https://medium.com/imagine-social-media/what-is-happening-on-tiktok-shadowbanning-and-censorship-6fba2c460524
[47] https://www.inverse.com/input/culture/tiktok-censored-banned-pole-dancer-phd-carolina-are
[48] https://www.forbes.com/sites/johnkoetsier/2021/12/27/top-10-most-downloaded-apps-and-games-of-2021-tiktok-telegram-big-winners/?sh=4fd235323a1f
[49] https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you
[50] https://dev.to/mage_ai/how-does-tiktok-use-machine-learning-5b7i



Despite this multifaceted approach to analyzing content, many users who had experienced shadowbanning on Instagram encountered a similar phenomenon on TikTok.

In the aforementioned article, Dr. Are spoke on the difficulties she had with maintaining a TikTok account. Users misflagged her content by reporting videos for "adult nudity and sexual activity," despite all of her posts consisting of her dancing clothed. This caused an "automatic suspension," which prevented Dr. Are from posting new content. Because of these repeated reports, she was also taken off the creator fund (effectively demonetizing her videos). "You have zero communication with the platform to understand what's going on as a user," she explained. "Other pole dancers and instructors said the same — some were even deleted."

Interestingly, Dr. Are's posting rights were reinstated once the *Input* article's author approached TikTok about her suspension. This also occurred when Eva Fog Noer, who posts content that raises awareness about child sex trafficking, was banned from the site for "violating community guidelines." After a reporter reached out to the platform about this unjust ban, Noer was able to post again.

In response to surveillance from both the algorithm and users intent on abusing the report system, many creators have adapted by using "algospeak" to convey information related to sexual health and

sexuality.[51] “Seggs” educators such as Madeline Gregg used mispelled words and euphemisms to talk about sexual health topics. They also add content warnings and disclaimers at the beginnings of their videos and carefully “watch their language and even their movements.”[52] "If you use the wrong gesture with your hand, say just one wrong word, it's enough to trigger [deletion]," said Dr. Alicia Jeffrey-Thomas, a pelvic floor PT. “It's mysterious to me...it's just frustrating.”

In late 2022, TikTok released a response[53]on how the platform would further enforce content deemed mature by the algorithm. “We've always had strict policies prohibiting nudity, sexual activity, and sexually explicit content, including content that directs to adult websites or apps,” the statement on Newsroom read. “Over 40,000 dedicated trust and safety professionals work to develop and enforce these policies and build processes and technologies to detect, remove or restrict violative content at scale.” In addition to overtly mature content, the platform asserted its enforcement of “borderline” mature content, which would be restricted from appearing on the FYP. Many users believe that this statement is a response to the rise in algospeak, which gained notable media attention just weeks before.[54][55]

I didn’t begin utilizing Lips’ TikTok page until the aforementioned strikes against the Instagram account. After a discussion with my internship director, it was decided that Lips should shift its advertising away from Instagram and to TikTok.

The first task was the curate the algorithm to suit Lips’ niches: I liked posts from creators whose content centered sex positivity and uplifting marginalized groups and followed relevant hashtags such as #queerart and #swer. Within minutes of doing this, the account’s FYP caught on to these niches and

[51] https://blog.hootsuite.com/social-media-definitions/algospeak/#:~:text=Algospeak%2C%20or%20algorithm%20speak%2C%20refers,or%20%F0%9F%8C%BD%20instead%20of%20porn
[52] https://mashable.com/article/tiktok-sex-education-content-removal
[53] https://newsroom.tiktok.com/en-us/strengthening-enforcement-of-sexually-suggestive-content
[54] https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/
[55] https://www.forbes.com/sites/anthonytellez/2023/01/31/mascara-unalive-corn-what-common-social-media-algospeak-words-actually-mean/?sh=531d913e2a08

@lip.social reached "the right side of TikTok." The feed was now filled with creators who would be interested in a platform like Lips.

When interacting with these creators and consuming their content, I found similar patterns to what has already been established in this case study. Many users utilize algospeak, replacing words like "kink" with "qink" (pronounced "kwink") , "sex" with "smex," and "sex work" with "spicy work." However, even when users incorporate algospeak into their language and tags, many still struggle to maintain engagement and avoid strikes against their accounts.

Hope Vicious (@thehopevicious), an outspoken SWer of 13 years, gets less than 1000 views on 87% of her posts, despite having over 43k followers. Like others, Vicious sees this as a form of shadowbanning. The hashtag #shadowban has over 4.7 *billion* views on the platform, and there are countless videos of users disdained by lack of engagement. @lips.social does yet have enough posts for me to compile adequate data on our own account's view suppression, but many of the users who are outspoken about shadowbans on this platform as queer and/or sex-positive creators.

Case Study: Lips.social

Lips was created as an antithesis to the censorship-savvy algorithms of sites like Instagram and Tumblr. A statistic often highlighted by its developers is that one year post-FOSTA-SESTA, a CHEQ survey found that 73% of content from LGBTQ+ sites such as *them* and *The Advocate* were "blacklisted for advertisers" on several social media sites.[56] According to its funding page, Lips was envisioned as a space for queer creators, including SWers and erotic artists, to safely showcase their work.[57]

Currently, 38.5k users and over 2500 active creators are on Lips.social. Many of these users are openly queer and work in the sex industry. To foster this space, Lips founder Annie Brown has based the

[56] https://www.advocate.com/media/2019/9/16/over-73-percent-lgbtq-content-online-flagged-inappropriate
[57] https://wefunder.com/lips/

platform in the blockchain. Payment processes, link-sharing, and other exchanges are decentralized and encrypted to keep all users safe. Lips is also not available on the App Store to prevent "Apple's monopolistic and conservative forces" from censoring the platform from potential users.[58]

Lips.social is explicitly meant for mature audiences, and users must fill out creator applications to fully access the site and post their own content. According to a statement released on the @lips_zine IG page: "Lips contain explicit nudity therefore you must be at least 17 years old to browse and at least 18 years old to post nudity. We decided to target this age range (17-35) most intentionally because these are the years that often tend to be most challenging on our mental health yet also incredibly formative for identity, confidence and sexuality. 17 is also the age restriction for adult apps in the Apple/Android stores."

These safeguards prevent minors and users who do not wish to see erotic content from using the site. For those who do want to use Lips but prefer less explicit content, the platform is very customizable thanks to its user-based data labeling "tags" system. By "tagging" posts, users are able to contextualize both their own and others' content. This makes certain posts, such as those tagged under #queerart, easier for others to find. While Tumblr and IG have similar systems, Lips' tags are also used by human moderators to efficiently investigate content with user context added to it. For instance, a moderator may feel inclined to delete a post with full nudity shown, but since the post is tagged #bodypositivity, this informs the mod of the user's intent.

Lips has a very small team of human content mods. There are plans to incorporate contextual-based AI into the platform, but currently, moderation is done manually.

As a new intern for Lips, I was introduced to the administrative content moderation page early on. This page was my new office– the space where I could look at flagged posts and creator applications. To post on Lips.social as a creator, one must apply with a description of the content they want to contribute

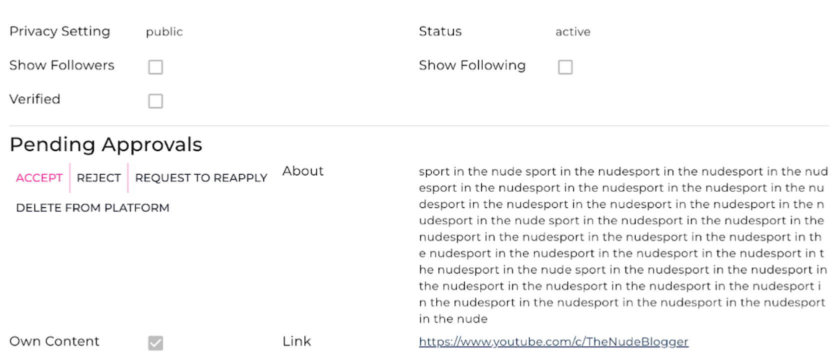
[58] https://bloggeronpole.com/2021/10/interview-with-social-media-platform-lips/

to the site and example posts. As one of the few active moderators, I found that there were over 1700 pending creator applications waiting to be looked at. For each one, I had the option to accept, decline, or request that users reapply based on what they share.

My first stage of content moderation can best be described as “taking out the trash.” I started with applications that I could easily decline – obvious spam and bot accounts with uncredited content (see image 6), users trying to promote harmful content, men looking to ogle at the site’s mainly femme and LGBTQ+ user base, etc. The last group ended up being the most common. Although most of these applications technically don’t violate any guidelines, I am told to reject them. These users would most likely spam and harass others on the site, and wouldn’t contribute quality art; thus, declining them is a preventative measure. I imagined that it would be difficult to program an AI mod that understands that.

Occasionally, I encountered applications that I was morally unsure about. For example, there were three accounts from sex workers who live in countries that have criminalized the profession. In their applications, they expressed that they wanted to post on Lips because all other SW sites were blocked where they’re from. Should I let them post on Lips, even though it is technically illegal in their countries?

I scheduled my weekly meeting with my internship director to address these applications. I expect her to tell me to reject them, but the opposite happens: “Since they don’t have anywhere else to post, let Lips be their safe haven,” Annie says. “That’s why I made this site – to give censored creators a space to post freely.”

6. 

By the time the second week of my internship rolls around, I've already cut the number of unreviewed creator applications in half. Admittedly, I was hasty on some of my decisions last week – but starting now, I was more careful when analyzing each application. Most of the applications I saw this week were ones that I either accept or leave to the side out of uncertainty.

Although Lips is very accepting of sex work and depictions of nudity compared to other social media, anything that clearly depicts real-life sexual acts goes against community guidelines. (Artistic depictions of acts, which will be discussed later, fall into a gray zone that requires me to assess their subjective "artistic value" more critically.) For SWers who add such content to their applications, I reached out to them and ask that they reapply using different example posts. The same process occurred when I came across users who put personal information in their bios, to protect both their safety and others' (ex: a SWer adding their personal phone number for chats). I was more reluctant to decline the applications of SWers or those in the LGBTQ+ community, since these are the groups that Lips caters to.

On week three, I decided to shift my focus to monitoring Lips' main page. The posts on Lips.social appear as squares of visuals sorted in reverse chronological order, reminiscent of both Instagram's grid and Twitter's early timeline format. Everything is sorted chronologically, with the most recent posts at the top. This work was a lot simpler than the creator applications process – whenever I saw something that violated Lips' guidelines, all I had to do was copy the post's link and paste it to my admin account's "moderation" page. Even though I have the ability to immediately delete posts, I began by sending them to Annie or other members of the Lips development team to look at.

As mentioned earlier, I was advised to flag posts with overt sexual acts. If there was an account with multiple violations, I sent a message to the user and asked that they reconsider the things they post. For example, I reached out to an account that was posting realistic AI generations of models. Not only were these images very sexually explicit, but I found it concerning that they seemed to be based on real people. After I expressed these concerns, the user responded saying that they didn't know that many AI platforms use real people to generate images, and they switched to using animated AI generators.

I continued these tasks for subsequent weeks, regularly checking both the Lips creator applications page and the Lips.social feed. At the same time, I observed meetings regarding the development of Reliabl.ai, a community-driven data management program that will be implemented into Lips' content moderation system in the near future. This program centers "bottom-up of participatory" practices that promise an inclusive social media setting.[59] With Reliabl, users could annotate their own posts with contextual information, which will be conveyed to both human moderators and the algorithm. As the program's website states, "[t]his approach helps to shift power away from centralized groups of homogeneous individuals, and towards more diverse communities."

What this entails specifically is still in progress. Not a meeting goes by without developing team members reiterated the importance of centering Lips' community. However, concerns have been raised about giving too much categorization power to users, as some may post dangerous content and bypass the platform's community guidelines by labeling their posts as innocuous. Similar security risks are assessed in these meetings as Reliabl's annotation system is fleshed out.

[59] https://reliabl.ai/

III. Lips, Unpursed: Discussion of Case Studies & Potential Solutions to Over-Censorship of Consensual Sexuality

As I finish my internship with Lips, the answers to the questions initially presented in this paper have become blurrier rather than clearer. When does a content moderator's power to censor encroach on another's free speech? When does the erotic become pornographic? When (if ever) will AI understand the intricacies of sexual identity on the Internet? There were many times when I sought guidance from my internship mentor or reached out to Lips users for elaboration on what they've posted. The very existence of context-based nuances in content moderation is why centering the marginalized communities who can shed light on these nuances is so vital.

While curated algorithms (particularly TikTok's FYP) continue to draw in large online audiences, there have been several reported incidents of these automated systems possessing biases against certain users, particular those whose content is related to queer identity, erotic art, sexual health, or other "borderline sexual" topics. After the passing of FOSTA-SESTA, such incidents increased. We have seen "shadowban" enter the dictionary of users across social media, regardless of how different algorithms function in changing user engagement. Instagram and TikTok's shadowban presents itself as the removal of a flagged user's content from the explore feed (causing a notable loss in views/likes), whereas being shadowbanned on Tumblr means having one's content mislabeled as NSFW and removed from appearing under certain hashtags.

In order to ensure that censorship efforts are well-informed and efficient, large social media platforms need to address how overgeneralized censorship methods can marginalize communities who value sexual expression. Additionally, it is imperative that the representative staff of social media sites actively communicate with the users of their platforms about censorship problems.

A common theme amongst the case studies featured in this paper is division: social division between users and administrative representatives, along with functional divisions between human

moderators and automated algorithms. Regarding the former, Tumblr's staff seem to be the most transparent with users, responding quickly to concerns expressed on the site. TikTok, in contrast, does not respond to users' appeals to suspensions/bans unless a secondary user reaches out to them. Such division can cause a loss in user trust regarding a platform's moderation ethics – this is seen in the backlash META has received over the years from vague and contradictory community guidelines.

Human moderators and automated algorithms are often treated as separate entities, rather than two parts of a comprehensive moderation system. This causes additional ethical issues related to oversight from both entities, from AI algorithms misflagging posts based on biased criteria to human error.

HITL systems and community data labeling have the potential to improve the state of moderation, with a people-focused approach in mind. Human mods can check the posts being flagged by AI, and vice-versa. The use of community-monitored tag categories offers another level of agency for both users and human mods. As shown by the popularity of algospeak on TikTok, slang on the Internet is constantly evolving and people are historically much better at understanding the nuances of language, compared to automation. Still, AI algorithms can be used as an initial step in efficiently identifying and categorizing data. For example, an AI program could filter flagged posts that are potentially disturbing, which gives human mods a content warning before they review the posts. Similarly, both user-driven annotation tools and AI can be used to spot spam posts before mods spend time investigating them.

Overall, practices that promote transparency and participatory collaboration must be incorporated into content moderation to alleviate biases against consensual expressions of sexuality.

IV. Reading Lips: Bibliography (APA7)

1. Hootsuite Inc. *2022 digital trends report*. Digital Trends - Digital Marketing Trends 2022. https://www.hootsuite.com/resources/digital-trends
2. Ali, S. S. (2017, February 5). *Human trafficking increased in 2016, organization reports*. NBCNews. https://www.nbcnews.com/news/us-news/human-trafficking-increased-2016-organization-reports-n717026
3. Couch, R. (2014, July 25). *70 percent of child sex trafficking victims are sold online: Study*. HuffPost. https://www.huffpost.com/entry/sex-trafficking-in-the-us_n_5621481
4. Murray, Z. (2023, June 30). *[Student Essay] How US law fosta-sesta "inadvertently" censored sex workers' social media usage*. SciencesPo. https://www.sciencespo.fr/public/chaire-numerique/en/2023/06/30/student-essay-how-us-law-fosta-sesta-inadvertently-censored-sex-workers-social-media-usage/
5. Romano, A. (2018, April 13). *A new law intended to curb sex trafficking threatens the future of the internet as we know it*. Vox. https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom
6. Wagner, A. *Text - H.R.1865 - 115th congress (2017-2018): Allow states and victims ...* Congress.gov. https://www.congress.gov/bill/115th-congress/house-bill/1865/text?overview=closed
7. Moraff, L. (2023, February 24). *How online censorship harms sex workers and LGBTQ communities: ACLU*. American Civil Liberties Union. https://www.aclu.org/news/civil-liberties/how-online-censorship-harms-sex-workers-and-lgbtq-communities
8. Tung, L. (2020, July 10). *Fosta-Sesta was supposed to thwart sex trafficking. Instead, it's sparked a movement*. WHYY. https://whyy.org/segments/fosta-sesta-was-supposed-to-thwart-sex-trafficking-instead-its-sparked-a-movement/
9. Robertson, A. (2021, June 24). *Internet sex trafficking law fosta-sesta is almost never used, says government report*. The Verge. https://www.theverge.com/2021/6/24/22546984/fosta-sesta-section-230-carveout-gao-report-prosecutions
10. Mohan, P. (2022, April 15). *Four years after Sesta/FOSTA, a new bill investigates its harm*. Fast Company. https://www.fastcompany.com/90741323/four-years-after-sesta-fosta-a-new-bill-investigates-its-harm
11. Mauro, G., & Schellmann, H. (2023, February 8). *"there is no standard": Investigation finds AI algorithms objectify women's bodies*. The Guardian. https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies
12. Dacon, J. (2023, September 18). *Are you worthy of my trust?: A socioethical perspective on the impacts of trustworthy AI systems on the Environment and Human Society*. arXiv.org. https://arxiv.org/abs/2309.09450
13. ActiveFence. (n.d.). *Spectrum labs: Trust and safety strategies*. Spectrum Labs | Trust and Safety Strategies. https://www.spectrumlabsai.com/trust-and-safety/#:~:text=Contact%20Us%20Today-,What%20is%20Trust%20and%20Safety%3F,that%20are%20outside%20community%20guidelines
14. Mage. (2022, February 16). *How does tiktok use machine learning?* DEV Community. https://dev.to/mage_ai/how-does-tiktok-use-machine-learning-5b7i

15. Wang, G. (2019, October 20). *Humans in the loop: The design of interactive AI Systems*. Stanford HAI. https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems
16. Cole, S. (2018, July 31). *Where did the concept of "Shadow banning" come from?* VICE. https://www.vice.com/en/article/a3q744/where-did-shadow-banning-come-from-trump-republicans-shadowbanned
17. Shvartsman, D. (2023, October 26). *Facebook: The leading social platform of our times*. Investing.com. https://www.investing.com/academy/statistics/facebook-meta-facts/#:~:text=More%20than%2077%25%20of%20Internet,by%202%2C203%25%20over%20ten%20years
18. Fisher, M. (2018, December 27). *Inside facebook's secret rulebook for global political speech*. The New York Times. https://www.nytimes.com/2018/12/27/world/facebook-moderators.html
19. Jee, C. (2020, June 8). *Facebook needs 30,000 of its own content moderators, says a new report*. MIT Technology Review. https://www.technologyreview.com/2020/06/08/1002894/facebook-needs-30000-of-its-own-content-moderators-says-a-new-report/
20. Barrett, P. M. (2020, June). *Who Moderates the Social Media Giants? A Call to End Outsourcing*. NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/tech-content-moderation-june-2020
21. Marr, B. (2021, December 10). *The amazing ways instagram uses big data and Artificial Intelligence*. Forbes. https://www.forbes.com/sites/bernardmarr/2018/03/16/the-amazing-ways-instagram-uses-big-data-and-artificial-intelligence/?sh=1037bac95ca6
22. Emerging Technology from the arXiv. (2020, April 2). *Data-mining 100 million Instagram photos reveals global clothing patterns*. MIT Technology Review. https://www.technologyreview.com/2017/06/15/105762/data-mining-100-million-instagram-photos-reveals-global-clothing-patterns/
23. *Instagram Community Guidelines*. Instagram Help center. (n.d.). https://help.instagram.com/477434105621119
24. *Sexual solicitation*. META Transparency Center. (n.d.). https://transparency.fb.com/policies/community-standards/sexual-solicitation/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fsexual_solicitation
25. Canales, K. (2021, March 25). *Mark Zuckerberg said content moderation requires "nuances" that consider the intent behind a post, but also highlighted Facebook's reliance on ai to do that job*. Business Insider. https://www.businessinsider.com/zuckerberg-nuances-content-moderation-ai-misinformation-hearing-2021-3
26. Petty, C. (2022, November 16). *The Naked Truth: Meta's Censorship of Sexual Health Information and Advocating to Big Tech for Change*. RNW Media. https://www.ohchr.org/sites/default/files/documents/issues/digitalage/cfis/tech-standards/subm-standard-setting-digital-space-new-technologies-csos-choice-rnw-media-3-input-part-2.pdf
27. Iqbal, N. (2020, August 9). *Instagram "censorship" of black model's photo reignites claims of race bias*. The Guardian. https://www.theguardian.com/technology/2020/aug/09/instagrams-censorship-of-black-models-photo-shoot-reignites-claims-of-race-bias-nyome-nicholas-williams
28. PLEASERS SERIES. (2022, September 26). *Sex face: Censorship and it's impact on Sex work*. YouTube. https://www.youtube.com/watch?v=ZO1CsWRXEcE

29. Ohene, A. (2022, May 16). *School of sex ed's Instagram account censored and deactivated by algorithms*. School of Sexuality Education. https://schoolofsexed.org/blog-articles/2020/09/17
30. Tabahriti, S. (2022, August 28). *Mark Zuckerberg says there is no "Shadow banning" on Facebook but admits there are "millions of mistakes."* Business Insider. https://www.businessinsider.com/mark-zuckerberg-no-shadow-ban-facebook-but-mistakes-are-made-2022-8
31. Mosseri, A. [@mosseri] (2023, June). *How the "Algorithm" Works.* [Instagram Post] https://www.instagram.com/reel/Cs6gh_NgPF0/?utm_source=ig_embed&ig_rid=daac66e5-112d-4b18-bd78-0e271570bcd2
32. Haimson, O. L., Dame-Griff, A., Capello, E., & Richter, Z. (2019, October 18). *Full article: Tumblr was a trans technology: The meaning, importance, history, and future of trans technologies*. Taylor & Francis Online. https://www.tandfonline.com/doi/full/10.1080/14680777.2019.1678505
33. Tumblr. (n.d.). *About Tumblr*. Tumblr. https://about.tumblr.com/
34. Tumblr. (n.d.). *Content moderation on tumblr – help center*. Tumblr Help Center. https://help.tumblr.com/hc/en-us/articles/360011799473-Content-moderation-on-Tumblr
35. Perez, S. (2017, June 24). *Tumblr says it fixed the "Safe Mode" glitch that hid innocent posts, including LGBTQ+ content*. TechCrunch. https://techcrunch.com/2017/06/24/tumblr-says-it-fixed-the-safe-mode-glitch-that-hid-innocent-posts-including-lgbtq-content/
36. Staff. (2017, June 23). *Changes to self-marked blogs*. Tumblr. https://staff.tumblr.com/post/162178688374/safe-mode-update
37. Perez, S. (2017a, June 20). *Tumblr rolls out new content filtering tools with launch of "safe mode."* TechCrunch. https://techcrunch.com/2017/06/20/tumblr-rolls-out-new-content-filtering-tools-with-launch-of-safe-mode/
38. Shannon, J. (2018, November 21). *Discovery of child pornography leads to Tumblr's removal from Apple's App Store*. USA Today. https://www.usatoday.com/story/tech/news/2018/11/20/tumblr-ios-app-store-apple-child-pornography/2073740002/
39. Fahy, C. (2022, November 2). *Tumblr says clothing is optional again*. The New York Times. https://www.nytimes.com/2022/11/02/technology/tumblr-nudity-explicit-posts.html
40. Jackson, G. (2018, December 13). *Tumblr porn ban leaves artists and fans seeking new platforms*. Kotaku. https://kotaku.com/tumblr-porn-ban-leaves-artists-and-fans-seeking-new-pla-1831056412
41. Metz, R. (2019, January 2). *We tested Tumblr's ban on porn. it needs work | CNN business*. CNN. https://www.cnn.com/2019/01/02/tech/ai-porn-moderation/index.html
42. Asher [@gachabasta] (2021, May 4). *Tumblr after your ban on "adult content" you removed all of my top surgery recovery photos citing that my nipples were "female-presenting"/So you took down my photos and actively misgendered me in the process./I can only say: who do you think you are* [Tweet]. https://twitter.com/gachabasta/status/1389758728056950789?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1389758728056950789%7Ctwgr%5Eb22a6965487925b4ccf96d5b03a47c3736d34d34%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fwww.vice.com%2Fen%2Farticle%2F93yyp8%2Ftumblr-is-trying-to-win-back-the-queer-audience-it-drove-off

43. Medlen, D. R. (2022, September 28). *Is NSFW back on Tumblr? community labels on Tumblr explained*. The Mary Sue. https://www.themarysue.com/is-nsfw-back-on-tumblr-community-labels-on-tumblr-explained/
44. Angelltheninth. (2023, January 27). *Community Labels Are Killing Fandoms and Creators*. Tumblr. https://www.tumblr.com/angelltheninth/707593013502836736/community-labels-are-killing-fandoms-and-creators?source=share
45. Lee, D. (2018, August 2). *The popular Musical.ly app has been rebranded as TikTok*. The Verge. https://www.theverge.com/2018/8/2/17644260/musically-rebrand-tiktok-bytedance-douyin
46. Biaso, M. L. (2021, August 20). *What is happening on TikTok: Shadowbanning and censorship*. Medium. https://medium.com/imagine-social-media/what-is-happening-on-tiktok-shadowbanning-and-censorship-6fba2c460524
47. Stokel-Walker, C. (2021, February 22). *TikTok censored a pole-dancing Phd who studies how Social Media Silences Women*. Input. https://www.inverse.com/input/culture/tiktok-censored-banned-pole-dancer-phd-carolina-are
48. TikTok. (2022, December 30). *Strengthening enforcement of sexually suggestive content*. Newsroom. https://newsroom.tiktok.com/en-us/strengthening-enforcement-of-sexually-suggestive-content
49. Woodward, M. (2023, September 20). *Tiktok User Statistics 2023: Everything you need to know*. SearchLogistics. https://www.searchlogistics.com/learn/statistics/tiktok-user-statistics/
50. Shead, S. (2020, November 13). *TikTok is luring Facebook moderators to fill new trust and Safety Hubs*. CNBC. https://www.cnbc.com/2020/11/12/tiktok-luring-facebook-content-moderators.html
51. Koetsier, J. (2022, November 9). *Top 10 most downloaded apps and games of 2021: Tiktok, telegram big winners*. Forbes. https://www.forbes.com/sites/johnkoetsier/2021/12/27/top-10-most-downloaded-apps-and-games-of-2021-tiktok-telegram-big-winners/?sh=4fd235323a1f
52. TikTok. (2020, June 18). *How Tiktok recommends videos #ForYou*. TikTok Newsroom. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you
53. Lorenz, T. (2022, April 11). *Internet "Algospeak" is changing our language in real time, from "nip nops" to "le dollar bean."* The Washington Post. https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/
54. Tellez, A. (2023, February 1). *"mascara," "unalive," "corn": What common social media algospeak words actually mean*. Forbes. https://www.forbes.com/sites/anthonytellez/2023/01/31/mascara-unalive-corn-what-common-social-media-algospeak-words-actually-mean/?sh=531d913e2a08
55. Mage. (2022a, February 16). *How does tiktok use machine learning?*. DEV Community. https://dev.to/mage_ai/how-does-tiktok-use-machine-learning-5b7i
56. Ogles, J. (2023, May 31). *Over 73 percent of LGBTQ content online flagged as "inappropriate."* The Advocate. https://www.advocate.com/media/2019/9/16/over-73-percent-lgbtq-content-online-flagged-inappropriate
57. Brown, A. (2021, November). *Lips: Instagram alternative built for women, the LGBTQ+ community, & their fans*. Wefunder. https://wefunder.com/lips/
58. Carolina. (2021, October 21). *Interview with Social Media Platform Lips*. Blogger On Pole. https://bloggeronpole.com/2021/10/interview-with-social-media-platform-lips/
59. https://reliabl.ai/

Additional Readings:

Dacon, J., Shomer, H., Crum-Dacon, S., & Tang, J. (2022, June 15). *Detecting harmful online conversational content towards LGBTQIA+ individuals*. arXiv.org. https://arxiv.org/abs/2207.10032

Blunt, D., & Wolf, A. (2021, October 17). *Erased - the impact of Fosta-Sesta and the removal of Backpage 2020*. Hacking//Hustling. https://hackinghustling.org/erased-the-impact-of-fosta-sesta-2020/

Chamberlain, L. (2019). *FOSTA: A Hostile Law with a Human Cost* . Fordham Law Review. https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5598&context=flr

Wijesiriwardena, S. (2019, June 24). *Private parts: Obscenity and censorship in the Digital age*. GenderIT.org. https://genderit.org/feminist-talk/private-parts-obscenity-and-censorship-digital-age

Knight, M. E. (2022, May 6). *#SeggsEd: Sex, safety, and censorship on TikTok*. SDSUnbound. https://digitallibrary.sdsu.edu/islandora/object/sdsu%3A200765

Ohlheiser, A. W. (2021, July 13). *Welcome to TikTok's endless cycle of censorship and mistakes*. MIT Technology Review. https://www.technologyreview.com/2021/07/13/1028401/tiktok-censorship-mistakes-glitches-apologies-endless-cycle/

McInerney, K., & Drage, E. (n.d.). *Feminism and Technology: The good robot podcast*. The Good Robot. https://www.thegoodrobot.co.uk/