A report from

cdt | Research

Shedding Light on Shadowbanning

Gabriel Nicholas

April 2022

The Center for Democracy & Technology (CDT) is a 25-year-old 501(c)3 nonpartisan nonprofit organization working to promote democratic values by shaping technology policy and architecture. The organisation is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

GABRIEL NICHOLAS

Research Fellow at the Center for Democracy & Technology.

A report from

cdt | Research

Shedding Light on Shadowbanning

Gabriel Nicholas

WITH CONTRIBUTIONS BY

Emma Llansó, Caitlin Vogus, Michal Luria, Dhanaraj Thakur, Samir Jain, Ari Goldberg, and Tim Hoagland.

ACKNOWLEDGEMENTS

We thank Sarita Schoenebeck and Kelley Cotter for their feedback on an earlier version of this report. We also sincerely thank the interviewees who shared their time and personal experiences on this topic, as well as industry workers who helped inform our analysis. All views in this report are those of CDT.

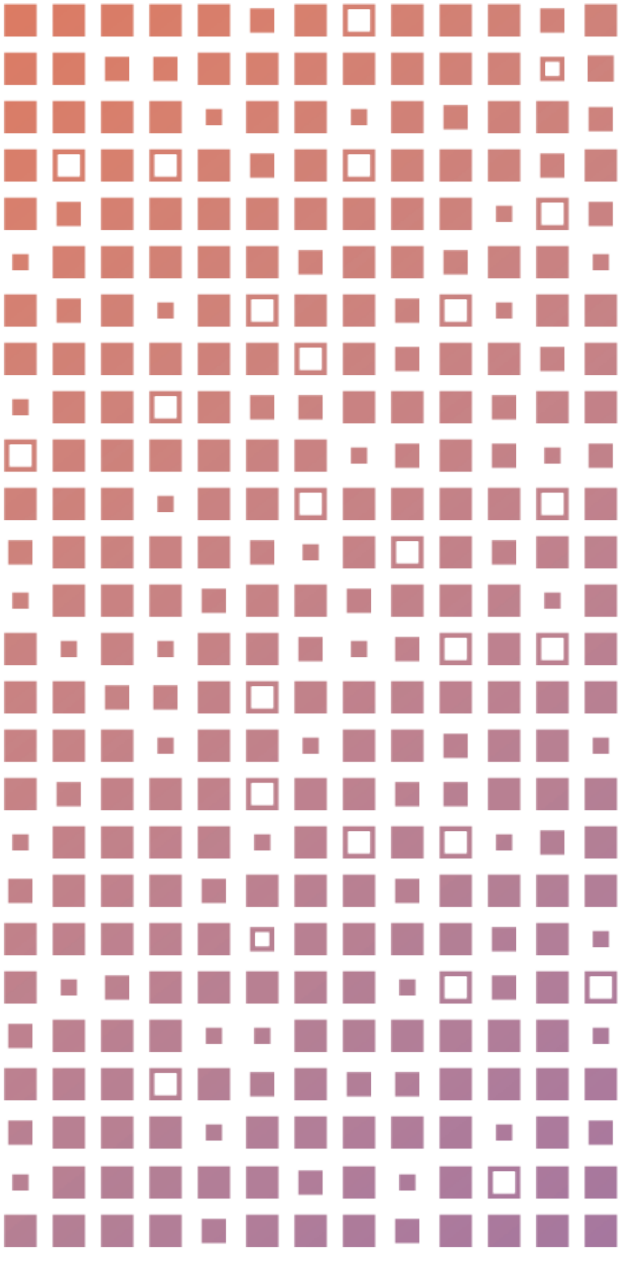
This work is made possible through a grant from the John S. and James L. Knight Foundation.

Suggested Citation: Nicholas, G. (2022). Shedding Light on Shadowbanning. Center for Democracy & Technology. https://cdt.org/insights/shedding-light-on-shadowbanning/

References in this report include original links as well as links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.

Contents

Introduction



For social media services, content moderation — the policies, practices and tools providers have in place to address online abuse — is their primary tool to defend users from everything from mis- and disinformation to hate speech and online extremism. As the list of potential abuses continues to grow, online service providers have commensurately built new systems to enforce their content policies, largely through removing or reducing the visibility of potentially abusive content (Douek, 2020).

But social media services don't always inform users when they are moderating their content: a user's comment may appear normally to themselves but be hidden to others; a user's handle may disappear in search; a user's original content may get ranked so low in a recommendation algorithm that, for all intents and purposes, it becomes undiscoverable. On the internet, people refer to this broad range of undisclosed content moderation actions as shadowbanning.

Recently, this term has moved front and center in debates about speech online. Black Lives Matter organizers claimed to be shadowbanned by TikTok after the killing of George Floyd (Gebel, 2020). President Trump tweeted the word in all caps, claiming that Twitter was "SHADOWBANNING" Republicans (Stack, 2018); he again decried social media services for shadowbanning him in his speech on January 6th (Trump, 2021). Legislators have also been paying attention to shadowbanning, with the term appearing in over a dozen state bills proposed by members of both major parties (Open States, n.d.). Politicians in Hungary and Poland have also publicly railed against shadowbanning and considered making it illegal (Szakacs, 2021; Walker, 2021).

Given all the public attention shadowbanning has received, there is surprisingly little published research on how shadowbanning impacts individuals, who it impacts, or what the word even means. There are a few possible reasons for this. Researchers may be hesitant to engage with the term's definitional ambiguity and heated political charge. Shadowbanning is also, by its opaque nature, difficult to study, especially at the level of confidence required for rigorous empirical research. Without social media companies' cooperation, it is difficult, if not impossible, for users or researchers to disambiguate when content

"You know, I feel a bit alone doing this research. I appreciate you contacting me." (Shadowbanned social media user, Interview, 2021)

"Shadowbanning is not a thing." Source - Adam Mosseri, CEO of Instagram (**Cook, 2020**)

is not getting engagement because it has been moderated from when it is simply uninteresting to users. The demand for information about shadowbanning so outstrips supply that the few who do research on the topic often find their inboxes flooded with pleas for help. One shadowbanning researcher described receiving so many anguished emails, "We could open a shrink's office."

Social media services are in general not forthcoming about when or whether they shadowban, denying it even in the face of leaks and "everyday algorithm auditing" (Shen et al., 2021) done by marginalized groups (e.g. Blunt et al., 2020; Human Rights Watch, 2021; Ravi, 2021; Salty, 2021). Services' secrecy around their shadowbanning practices has let misinformation about how they moderate speech thrive, as users lacking knowledge are more susceptible to arguments based on emotion (Moynihan, 1998). In debates about shadowbanning, social media companies take the defensive, users go on the offensive, and neither can meet in the middle because they don't share common language, knowledge, or goals.

The goal of this paper is to bridge the gap between social media companies, end users, and the broader public in how each understands shadowbanning in order to help social media companies better manage disclosures to users about content moderation. We aim to do this by critically examining three questions:

1. What is shadowbanning?
2. Who is affected by shadowbanning?
3. What larger effects does shadowbanning have?

To answer these questions, we took a mixed-methods approach in this research.[1] We conducted interviews with 36 participants — 13 were people who claimed to have been shadowbanned, 13 worked at social media services, and ten were members of academia and civil society who worked on issues relating to shadowbanning. We also commissioned an online, nationally representative survey of 1205 social media users in the U.S. to find out how many believed they had experienced shadowbanning and what their experiences were like. Additionally, we engaged in an expansive literature review, including news articles, public communications from social media companies, proposed laws, patents, and academic literature on shadowbanning and related phenomena.

1 See Appendix A for more details on these methods.

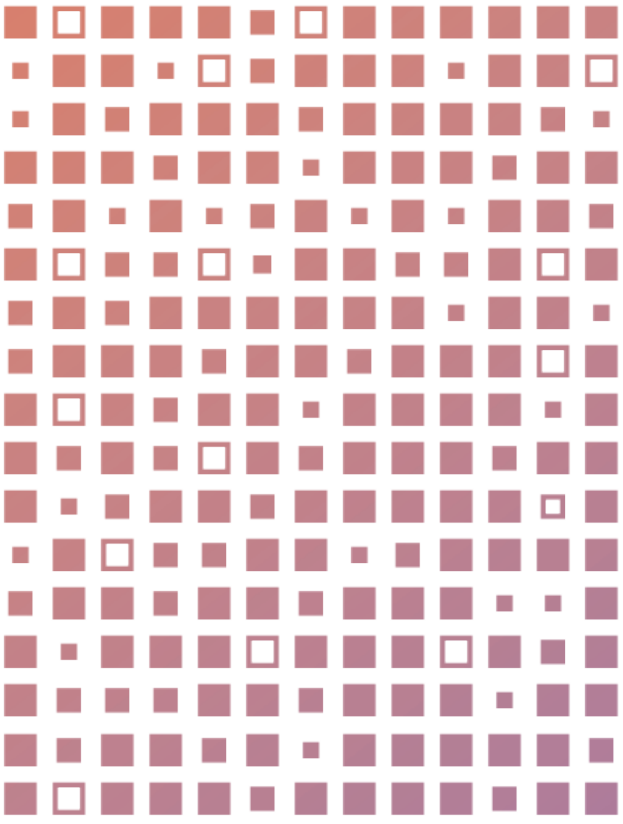
The first section of this paper will define the term shadowbanning, review specific shadowbanning practices in areas such as search, comments, and recommendations, and discuss the reasons social media companies give for why they engage in shadowbanning. The second section will look at which groups may be disproportionately affected by shadowbanning and describe how users diagnose and respond to their own shadowbanning. The third section will explore the consequences of shadowbanning on individuals, groups, and society at large.

The final section of this paper recommends three ways social media services can mitigate the harms of shadowbanning: sharply limiting the circumstances in which they shadowban, "shedding light" on shadowbanning by publishing their policies about when they shadowban and data about content or accounts subject to opaque content moderation, and creating mechanisms to allow researchers to learn more about the potential harms shadowbanning may cause.

What Is Shadowbanning?

"Twitter release me from twitter shadowban!!! I won't talk about [s*ckin] and [f*ckin] nomore. I promise that was 2020 behavior!"

Source - Cardi B's Twitter Account, January 3, 2021 (**Cardi B, 2021**)



Why use the term *shadowbanning*?

Before we begin to define it, it is worth explaining why we choose to use such a vague and controversial term as "shadowbanning" to talk about opaque content moderation practices instead of avoiding it entirely or creating our own term. In our interviews with members of civil society, many criticized the word as too imprecise, too politicized (particularly by conservatives), and too easily co-opted by people who have not actually been shadowbanned to stir up anti-tech outrage. More than having any specific, shared definition, "shadowban" is often a term of convenience employed by users who feel indignant that they are not getting the social media attention they believe they deserve.

Still, we decided to keep the term for three reasons. First, it is a word that people commonly use and recognize. Repeatedly in interviews, social media users were familiar with the term "shadowban" and used it themselves in a relatively consistent way to describe undisclosed content moderation actions. It is used by proposed legislation (e.g. Stop Social Media Censorship Act, 2021; Wisconsin Senate Bill 582, 2021), internet culture (e.g. Lorenz, 2017), and academics (e.g. Are, 2021; Myers West, 2018). Even online service providers, who often publicly and privately disparage the term for its vagueness, have used it in patents, including Facebook (Strauss et al., 2019) and Sony (Miyaki, 2021). Our survey also suggests that nearly half of US social media users are familiar with the term — 44% have either heard of the term shadowbanning (9%), are familiar with the practice (13%), or both (22%).

Second, the term "shadowbanning" effectively describes the phenomenon. The word "shadow" in particular calls to mind the practice's multiple levels of opacity. With shadowbanning, online services keep users "in the dark" about how their content is being moderated (Burrell, 2016; Eslami et al., 2015; Myers West, 2018). In another sense, users' content is sent to a "shadow realm" where no one else can see it — in the early 2010s, some also called the practice "hell banning" (Atwood, 2011; Rao, 2013). At a more meta level, service providers often do not admit to the practice (Cotter, 2021; Gadde & Beykpour, 2018; May, 2019), throwing these types of moderation actions further into the shadows.

▲ **Figure 1. Familiarity of the term "shadowbanning" among social media users in the U.S. (% of social media users in the U.S., n=1006).** Source - CDT National Survey of Social Media Users 2021

Third, avoiding the term "shadowbanning" and letting others define it comes with its own dangers. An overly narrow definition of the term could invalidate the lived experiences of those who have been shadowbanned, and potentially exacerbate the isolation and shame that comes with shadowbanning and other forms of content moderation, opaque and otherwise (Myers West, 2018). Yet an overly broad definition could fuel conspiracies and hostile anti-tech rhetoric (Barrett & Sims, 2021).

A working definition of shadowbanning

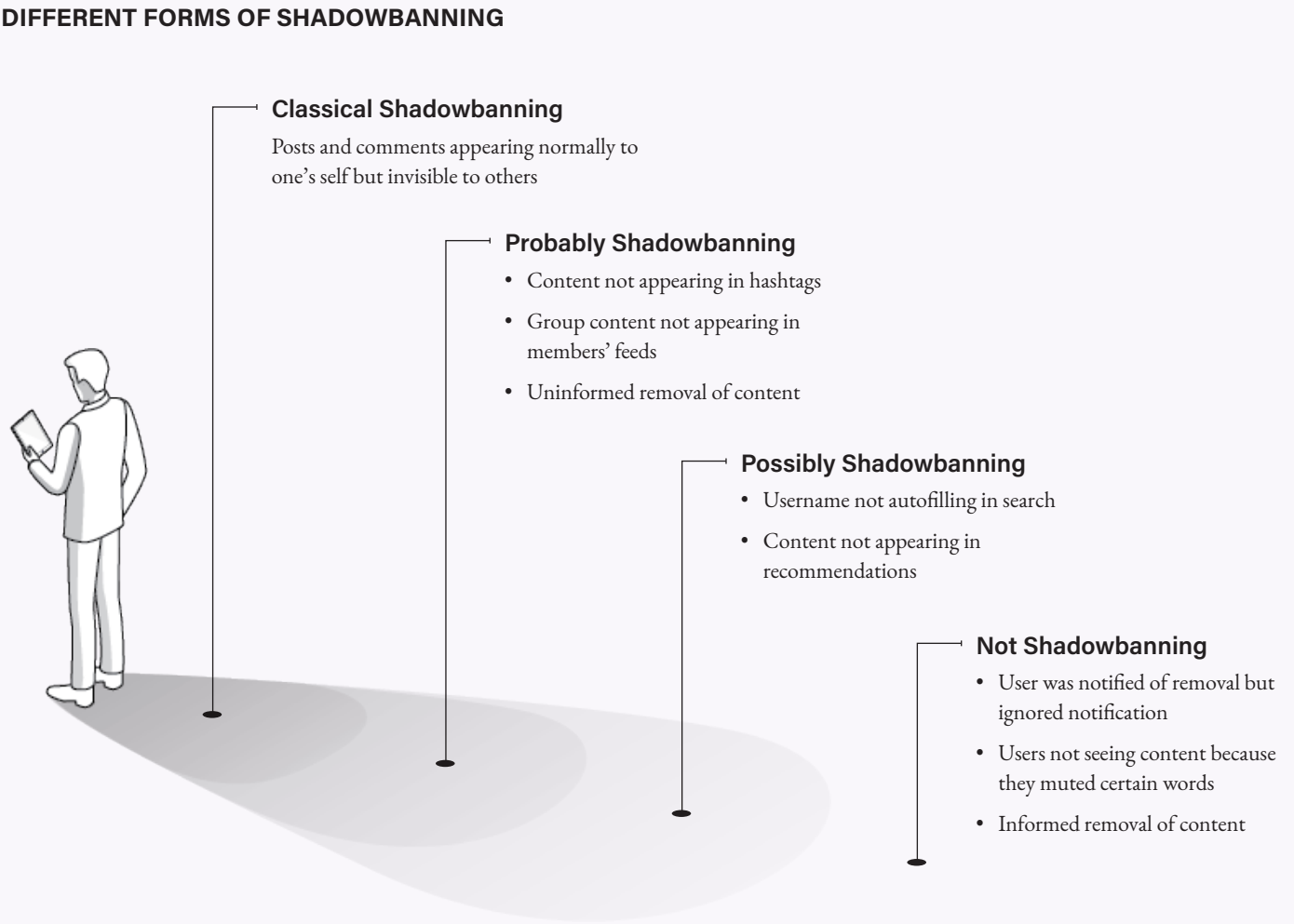
In our definition of "shadowbanning," we seek to take a descriptive view of how people actually use the term. This immediately raises problems because there is significant variation in how people use the term. At the core of most definitions is the concept of secretly hiding a user's posts while making it look to the user as if their posts are still publicly visible. This can be thought of as the "classical" definition of shadowbanning as it was used by social media users in the 2000s and early 2010s. However, social media have since grown in size, developed new features, and increasingly centered algorithms in their design. As a result, this classical definition fails to capture the many new ways users use the term more colloquially to describe other opaque forms of content moderation, such as search suggestion bans, not sending notifications, and hiding posts from recommendations.

A definition of shadowbanning that encompasses all existing definitions is so broad as to be useless. The broadest definitions of shadowbanning come from proposed state legislation. Sometimes, these definitions include all forms of content moderation and more, such as one Wisconsin bill that says (Wisconsin Senate Bill 582, 2021):

> "Shadow ban" means to limit or eliminate the exposure of a user, or content or material posted by a user, to other users of the social media Internet site through any means, regardless of whether the action is determined by an individual or an algorithm, and regardless of whether the action is readily apparent to a user.

To truly capture how people use the term, we cannot draw hard lines between what is and isn't shadowbanning. Instead, we define a center, of what is definitely shadowbanning, and an outer edge, of what could possibly be considered shadowbanning. At the center is undisclosed, undetectable content removal, the classical definition of shadowbanning. At the outer edge is any human or automated content moderation action, taken by the service, to remove or reduce the visibility of a user's content, of which the user is not informed.

One type of action that can possibly be considered shadowbanning is when service providers enforce their content policies with algorithmic deprioritization, usually in search or recommender systems (e.g. feeds, suggested content) (Gorwa et al., 2020). Some forms of algorithmic deprioritization are functionally equivalent to removal — for example, if a user's comment gets buried so deeply in other users' feeds that no one ever sees it. But not all algorithmic deprioritization is shadowbanning — recommendation algorithms also tend to deprioritize content that is less recent and receives less engagement from other users. For users, it is often impossible to tell whether their content is deprioritized because other users are genuinely not interested in it or because a service provider has taken action against it. Making this determination is especially difficult because there is no single "objective" ranking of all content against which a user can evaluate the placement of their post.



▲ **Figure 2. Different forms of shadowbanning.**

In this paper, we are focused on shadowbanning involving an enforcement action taken by the provider of an online service and therefore limit our definition to such cases. However, shadowbanning is not something inherently reserved for providers alone. On some services, community moderators can silently hide certain posts in groups or pages they run. Others allow users to mute keywords, either preventing themselves from seeing content that contains those words on their feed (e.g. Twitter Help Center, n.d.) or preventing others from seeing comments on their content that include certain flagged words (e.g. Instagram Help Center, n.d.; People for the Ethical Treatment of Animals v. Collins, 2021).

Content moderation is a complex process that occurs in a variety of phases; shadowbanning typically occurs in the enforcement and the education phases (**Kamara et al., 2021**).

Similarly, governments may also compel service providers to remove certain content without allowing them to tell the end user (Bloch-Wehba, forthcoming; Twitter Transparency Center, 2021), and while such gag orders can raise significant freedom of expression concerns, they are again outside the scope of this paper, which focuses on voluntary decision making by online service providers.

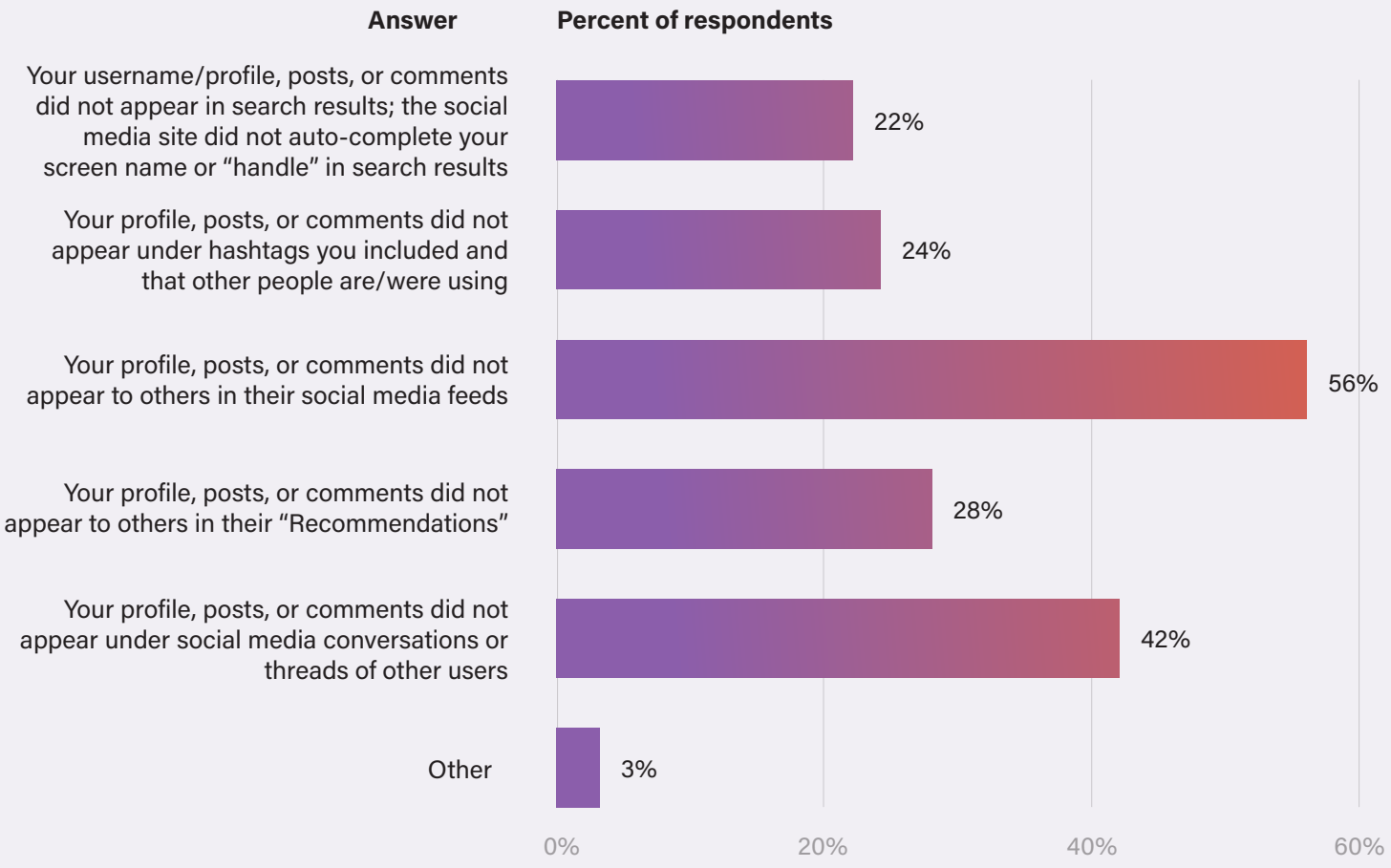
How do social media services implement shadowbanning?

By our definition, all major social media services engage in at least some moderation practice that falls along the spectrum of what we consider shadowbanning. Through interviews, leaked documentation, and our own testing on services, we uncovered a range of examples. There is an inherent limitation to the rigor of research that can be done on shadowbanning, since users often cannot confirm it themselves and service providers will typically not admit to it on the record. Still, there is so little research on shadowbanning that we believe it is worth sharing the practices we found and note what evidence we have for each to help improve our understanding of the problem.

In our survey, of the users who identified themselves as having been shadowbanned (n=274), most (56%) believed that their comments or posts were hidden from others' feeds and many (42%) believed their comments were not appearing in conversational threads. Fewer believed that their content was not appearing in search (22%), hashtags (24%), or recommendations (28%).

▼ **Figure 3. Different types of shadowbanning that social media users report experiencing (% of those who say they were shadowbanned, n=274).** Source - CDT National Survey of Social Media Users 2021

Thinking of the most recent time, to the best of your knowledge, what type of shadowbanning did you experience?



SEARCH

There are multiple ways a social media service may prevent content from appearing in search. The simplest way is to exclude a user or their posts from search results entirely. For example, according to Jack Dorsey's testimony before the U.S. House Committee on Energy and Commerce, Twitter excluded users they deemed likely to be abusive from search results in their "Latest" tab for a period of time (Hearing on "Twitter: Transparency and Accountability," 2018). Search bans are easily verifiable, since it only takes creating a new account and searching for one's own username.

Services can also enact suggestion bans, in which a user's name does not autofill in the search box when someone begins to type it in, making that user more difficult to find. Suggestion bans can be particularly difficult to detect because it is hard to disambiguate when a user's profile does not reach the popularity threshold the site has set to appear in search autofill versus when the service has intentionally prevented the user profile from appearing. Most famously, in 2018, a Vice article reported that Twitter was not autofilling the handles of many prominent Republicans in its search bar, including Republican National Committee chair Ronna McDaniel, former White House Chief of Staff Mark Meadows, and U.S. Reps. Jim Jordan (OH-4) and Matt Gaetz (FL-1) (Thompson, 2018). Twitter argued that this was a bug that was affecting far more accounts than it expected (Gadde & Beykpour, 2018), but Twitter continues the practice of not autofilling certain usernames, as we confirmed through interviews and the shadowban.eu tool.[2]

Some social media services also shadowban specific hashtags (Instagram, n.d.-g, n.d.-f). This sort of action is on the outer edge of what might be considered shadowbanning, since it does not target an individual's content but can affect communities that use those hashtags to gather. Sometimes, services shadowban hashtags to stop the spread of harmful ideas, even at times offering interventions (Gerrard, 2018). For instance, searching a pro-eating disorder hashtag like #thinspiration on Instagram produces a generic error page, while searching for #suicide gives the user a mental health warning and links to support materials (Instagram, n.d.-e). Hashtags only occasionally used for harmful content, such as #brain used for self-harm content, only show top posts from the hashtag instead of the top and most recent posts. Social media company workers we interviewed also suggested that groups of trolls or spammers sometimes organize

2 In September 2021, interviewees mentioned a few accounts that they believed had been shadowbanned. When we searched their handles on the shadowban.eu tool, we found it claimed that their usernames were not autofilling in search. When we went to type in their handles ourselves, we confirmed that they did not autofill until the entire username had been typed. The shadowban.eu tool closed down in January 2022.

through banal-seeming hashtags, or that attackers will flood hashtags they want to demote with pornography (e.g. Garcha, 2020). This may explain why on Instagram, innocent seeming hashtags such as #boho, #italiano, and #kansas show only top posts (Instagram, n.d.-a, n.d.-b, n.d.-c).

However, social media services have also shadowbanned hashtags to reduce the visibility of already marginalized communities. A report from the Australian Strategic Policy Institute, for example, found that TikTok was showing no search results for hashtags meaning "gay" in Russian, Estonian, and Bosnian, "transgender" in Arabic, and certain hashtags against leaders in Russia, Indonesia, and Thailand (Ryan et al., 2020, p. 5). As the report describes, "Users who are motivated enough can discover other posted videos using the hashtag, but only after posting their own video featuring the hashtag and then clicking through to the content there...Doing so directs the user to the correct page, which does exist, but is unavailable (shadow banned) from the platform's search results" (Ryan et al., 2020, p. 12).

COMMENTS AND POSTS

The paradigmatic form of shadowbanning is when a service provider "deliberately [makes] someone's content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster" (Gadde & Beykpour, 2018). Reddit is one of the few major services that has openly admitted to engaging in this practice (krispykrackers, 2015), though a Facebook patent for shadowbanning scammers in a marketplace suggests that Facebook Marketplace may use it as well. That patent uses the term "shadowban" itself — "The shadow-banned user may still compose and see their own posts in the marketplace, and may still compose and see their own outgoing messages, without indication that other users are not also seeing those posts and/or messages" (Strauss et al., 2019).

Instagram users have repeatedly described post content disappearing without their knowledge. In India for example, critics of the government have frequently claimed that their posted Stories have disappeared without them being informed (Ravi, 2021). Palestinian rights activists have made similar complaints about Instagram after the May 2021 Israeli-Palestinian clashes (7amleh, 2021; Human Rights Watch, 2021). Interviewees described other forms of comment and post shadowbanning, including having share buttons removed, being unable to tag people in posts, being unable to comment, receiving cryptic error messages when trying to post, and other users not getting notifications when someone posts, even if they signed up to be notified about that person's posts.

FEED AND RECOMMENDATIONS

On most major social media services, users discover new content through a feed (e.g. Facebook's Newsfeed, Twitter's Timeline, TikTok's For You page) or other ways of delivering recommendations such as YouTube's recommended video sections and Instagram's Explore page. Feeds and recommendations algorithmically curate content for users, and they create a fuzzy line between what content is not as visible due to opaque content moderation and what content is simply not capturing viewers' attention. There is no single "earned" priority that a user's content should receive in a recommendation algorithm — often, there isn't even a single shared ranking system for content, since service providers routinely test new versions of their recommendation algorithms (Soria, 2020). Some services, such as TikTok (TikTok, 2019), Facebook (Stepanov, 2021) and Instagram (Mosseri, 2021a), have released general guidance for what makes content more or less likely to appear on an individual's Explore or For You page. Instagram's Explore page also allows users to adjust the sensitivity of content they get recommended. All of these factors add to the complexity of determining when algorithmic downranking should or should not be considered shadowbanning.

Still, many service providers use opaque content moderation techniques that prevent a user's content from appearing on a feed without them being informed. Leaks from TikTok, for example, showed that moderators, at least at one point, could flag videos as "Not for feed" or "Not recommended" (Reuter & Köver, 2019). A study on Facebook Pages also strongly suggests that Facebook will temporarily prevent interest pages from appearing on the News Feed (Horten, 2021). Sometimes page moderators would be warned, but other times they would only find out when they saw a very sharp drop in their view metrics — in some cases over a 95% drop. Severely diminished appearances in users' feeds blur the line between what some social media employees refer to as "hard" versus "soft" content moderation (Gorwa et al., 2020).

Why do social media services shadowban?

In order to evaluate how, when, and whether social media services should engage in shadowbanning, it is important to understand the reasons service providers themselves give for the practice. We gathered these explanations from news coverage, academic papers, and, most importantly, background interviews with employees from service providers themselves.

Social media services face difficult trade-offs in their content moderation design choices because they face multiple competing incentives and have many stakeholders to manage, including posters, viewers, advertisers, shareholders, and governments (Burke et al., 2016). In some respects, services stand to gain from having opaque policies and practices, in particular through what Coyle & Weller call "constructive ambiguity" (2020). Stakeholders often have competing aims — for example, governments seeking a crackdown on allegedly illegal content versus advocates pushing back against overbroad removal of speech — so in order to achieve the appearance of consensus on policy, service providers may intentionally give stakeholders incomplete information about their goals and systems. So long as the service provider is trusted to make this tradeoff, it conveniently lets all stakeholders believe their interests are being met while also shielding service providers from public criticism and scrutiny (Pasquale, 2015).

"If an organization is not trusted, its automated decision procedures will likely also be distrusted."
Source - (**Coyle & Weller, 2020, p. 1433**)

In this section, we look at the justifications provided by service providers for two different levels of opacity in shadowbanning — reasons why they do not inform users that their content is being moderated, and reasons why they do not inform users of their policies for when and how they shadowban. (Almost all reasons they gave focus on the former, with the exception of one.) We also offer limitations and critiques for each justification.

SOCKPUPPET PREVENTION

On many social media services, when a user finds out that their content is being blocked or otherwise moderated, they will simply create a new, "sockpuppet" account and continue their behavior (Kumar et al., 2017; Solon, 2021). Users employ sockpuppet accounts to target abuse, coordinate disinformation, and spam, even after the service has taken action against the original account (Bohn, 2017; Oremus, 2019b; Stepanov, 2021). A badly behaved user may be slower to create a new account if they don't know that others cannot see the content posted from their existing account. Similarly, spammers that distribute their content through automated means may be slower to adjust their methods to avoid detection if they do not know they have been detected. Reddit has openly shared that it covertly hides some users' posts without informing them to avoid sock puppets (krispykrackers, 2015), and Facebook (Stepanov, 2021) and Instagram (Mosseri, 2021b) have admitted to hiding content in search and recommendations for similar reasons.

"I've personally talked to people in charge of large online communities – ones you probably participate in every day – and part of the reason those communities haven't broken down into utter chaos by now is because they secretly hellban and slowban their most problematic users"
Source - Jeff Atwood, founder of Stack Overflow (**Atwood, 2011**)

One challenge for service providers is that the automated detection systems they use to detect coordinated disinformation attacks from sockpuppet accounts often unintentionally flag online political activism (Starbird et al., 2019; Vargas et al., 2020). Enforcing sockpuppet bans with shadowbanning means that legitimate political speech may be hidden without any way for the affected users or the public to know that it is happening and provide feedback or criticism of service's actions.

REVERSE ENGINEERING AND GAMING

Workers at social media companies we interviewed also justified the use of shadowbanning as a way to prevent users from reverse engineering automated content detection systems. For example, social media company workers were concerned that users could use the receipt of a ban to determine which words a service did and did not consider a racial slur, by posting a series of racist terms under different accounts and determining which ones resulted in bans. This would then allow the user to express racist statements unmoderated, but in contravention of the intent behind the service's policy against racist slurs. Social media company workers expressed similar concerns about recommendations and feeds — they feared that if users could deduce too much information about what content gets suppressed by automated recommendation systems, they could exploit it. For one social media service, it took two years of internal discussion to even publish high level information about what went into their recommendation algorithm.

However, reverse engineering also helps users better understand a service's norms about what is and isn't acceptable behavior. If, for example, a user wants to create a transformative video work using a copyrighted music file, reverse engineering could allow that user to understand how many seconds of audio they could use without being moderated for copyright infringement, thus helping them better comply with the letter and the spirit of the rule (Ma & Kou, 2021). Shadowbans inhibit users' ability to use reverse engineering to learn and understand a service's rules.

OPERATIONAL AND DESIGN CHALLENGES OF COMMUNICATING CONTENT MODERATION TO USERS

Social media company workers repeatedly highlighted how difficult it is to communicate content moderation actions, especially in ways that all users can understand, regardless of digital literacy. Social media services need to give information to users about content moderation decisions in a way that they can understand and at a time when they can digest it. One worker interviewed mentioned that common ways of communicating information, such as pop ups, often get ignored. Social

media companies also do not want to overwhelm users with information, especially information that they may not care about. One company interviewed gave the example of how, as a form of content moderation, it turned off a feature that only some users utilized. When it informed users of the action, many were confused and annoyed, since not all users even knew the feature existed. All of these factors raise many real design challenges in informing users that their content has been moderated, leaving the de facto solution to be not informing them at all.

Social media companies also claim that variance in users' digital literacy makes communicating about algorithmic deamplification particularly challenging. Borderline content — content that approaches the border of breaking a platform's terms of service and is often particularly popular (Heldt, n.d.) — serves as a case in point. Twitch faced this problem for example, with the "hot tub meta" debate, when the "Just Chatting" live stream category on Twitch briefly became dominated by women streaming from hot tubs, wearing outfits that toed the line of Twitch's nudity and attire policy (Gonzalez, 2021). Social media company workers we interviewed said that they use automated methods to evaluate how likely a given piece of content is to violate their content policies and to proportionally limit the distribution of posts that approach the line. As they argued, this can be difficult to communicate to users since: a) users may not understand or care whether their content gets promoted by recommendation algorithms; b) the idea of turning down the likelihood of a piece of content appearing in someone else's recommendations is a difficult concept to communicate in a succinct way that users will engage with; and c) the service may not be able to extricate how much of a signal deboost a given piece of content receives is due to content moderation factors versus other factors, such as how many people are engaging with it. These latter two considerations, each an aspect of the challenge of explaining how machine learning models make decisions, also point to the overall difficulty in evaluating how well borderline content detection models perform at all (Shenkman et al., 2021).

Eventually, Twitch addressed the problem by creating a new category for hot tub streams (**Twitch, 2021**).

Importantly, all of these operational and design challenges only justify shadowbanning as a placeholder, not a permanent solution. No social media employee we spoke to portrayed figuring out how to properly inform users that their content has been moderated as an intractable problem, only as a difficult one that they have yet had the time and resources to solve.

PROTECTING HEALTH AND SAFETY

Social media company officials we interviewed argued that the drawbacks of shadowbanning could be worth it in situations where content moderation was used to protect the health and safety of themselves and others. One service provider, for example, said that, at some point (though they have since changed this practice), they shadowbanned pro-suicide content instead of removing it, in order to protect the posting user from emotional distress. Online services will also shadowban content in order to minimize its impact on other users. For example, services justify opaquely moderating pro-eating disorder content in order to stop people (mostly women) from "catching" anorexia by looking at images of other women (Gerrard, 2018). Similar language of "intellectual contagion" is used to talk about terrorist content (Baldauf et al., 2019; Midlarsky et al., 1980). In interviews, some social media company employees suggested that it was particularly important not to allow users to find ways to reverse engineer and circumvent health-related content moderation.

However, what the service provider deems "harmful" may not be obvious to users, who may not agree with the provider's definition or who may not experience harm in the same way that the provider predicts. The provider may also make mistakes in its assessment of what content meets its definition of "harmful." And, when they do not disclose what kind of content they shadowban, service providers can more easily moderate content that is innocuous or even societally beneficial, but that may be undesirable to the service provider for other reasons. For example, documents from TikTok leaked to *The Intercept* reveal that in 2019, TikTok moderators were instructed to hide from the For You recommendation page videos featuring people with "abnormal body shape," "ugly facial looks", "disabled people," "too many wrinkles," or with backgrounds in "slums, rural fields," and "dilapidated housing" (Biddle et al., 2020). Critics argued that TikTok wanted to push "ugly, poor, or disabled" users off the service, while a representative of TikTok said that the rules "represented an early blunt attempt at preventing bullying, but are no longer in place, and were already out of use when The Intercept obtained them" (Biddle et al., 2020). Shadowbanning makes it harder or even impossible to know whether service providers have made similarly blunt or misguided decisions about what content to moderate.

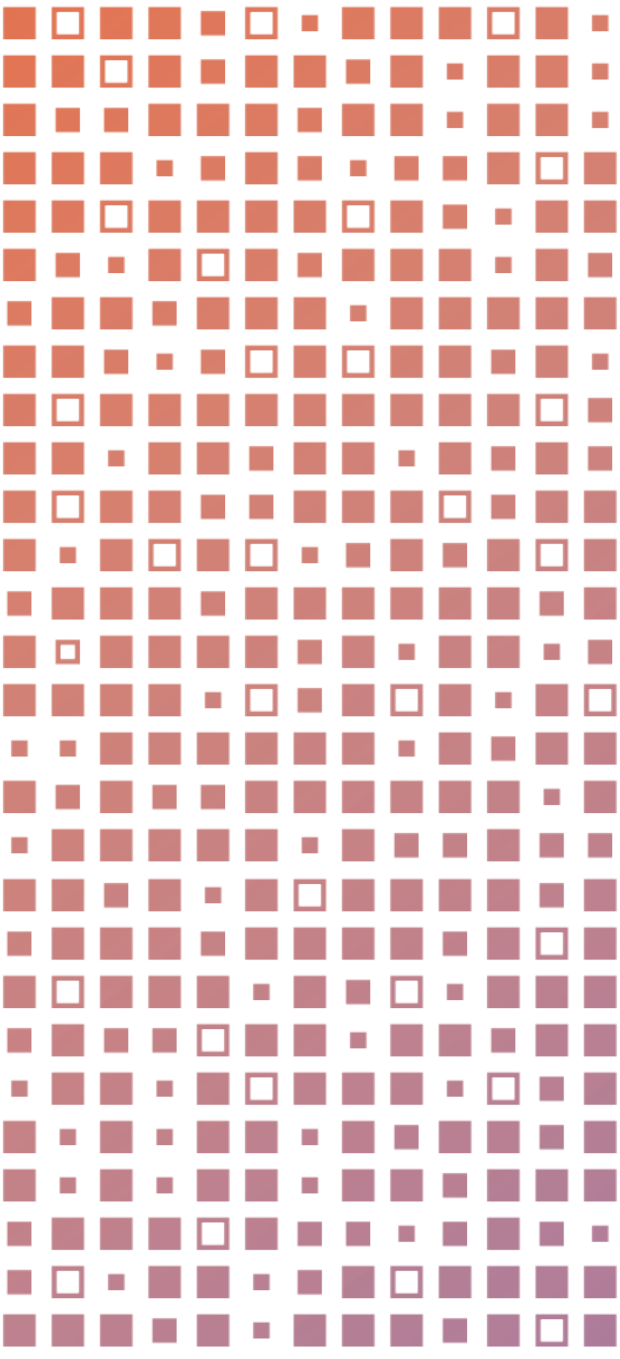
RESISTING AUTHORITARIAN GOVERNMENT DEMANDS

We heard only one argument for why social media services should not admit to engaging in the general practice of shadowbanning: social media companies claimed that they could not admit publicly that they shadowban content because, if they did, authoritarian governments would request them to shadowban. By not admitting they shadowban, social media companies argued that they could plausibly deny they had the technical capacity to do so.

This claim is difficult to independently verify since communications between governments and social media companies typically happen behind closed doors. Transparency reports, published by social media companies about government requests to moderate specific content, often provide only high-level information about what these interactions look like, mostly in numbers of removal requests and rate of different responses (Vogus & Llansó, 2021). Other sources of information, such as the Global Network Initiative public report on its assessment of member companies' adherence to the GNI Principles, provide additional detail about the nature of company-government interactions, in general, as well as case studies of specific (anonymized) interactions (Global Network Initiative, 2020). But on the whole, government-company interactions, especially those that involve extralegal government demands to implement technical restrictions on speech, occur out of the public view and are difficult to independently verify.

We do know, however, that governments across the world are aware of the concept of shadowbanning, including through their complaints that they suspect themselves to have been shadowbanned (Kaushika, 2019; Szakacs, 2021). Civil society and scholars in many parts of the world are already concerned that their governments may be asking or requiring service providers to shadowban (Lakier, 2021; Ravi, 2021). Additional transparency from providers about their policies around shadowbanning, and their policies and procedures for responding to government demands, would better enable journalists, advocates, and researchers to evaluate both governments' and companies' claims around opaque content restriction.

Who believes they are shadowbanned?



According to our survey, nearly one out of ten (9.2%) social media users believe they have been shadowbanned some time in the last year by a social media service. The perception of having been shadowbanned, however, isn't spread evenly across all populations. In this section, we present our survey and interview data about which groups are more likely to believe they have been shadowbanned and how they respond.

What groups believe they are affected by shadowbanning?

According to our survey, social media users who reported being shadowbanned were more often male, Hispanic, non-cis gendered, or Republican compared to the general population. More frequent social media users were also more likely to report having been shadowbanned, as were users who were either familiar with the term or practice of shadowbanning. It is important to note that these results do not reflect which social media users were *actually* shadowbanned, only which users perceived themselves to be shadowbanned. This limit is inherent to the difficulty or (in some cases) impossibility of verifying shadowbanning, and we do not have sufficient information to make any claims about the relationship between actual and perceived shadowbanning.

Still, other demographic and community research on content moderation aligns with these results and helps us better understand them. For example, research has found that Republican social media users more frequently report having their content removed by service providers, largely for being "offensive or allegedly so, misinformation, Covid-related, adult, or hate speech" (Haimson et al., 2021). Our survey similarly found that Republicans were more likely than Democrats and Independents to believe they were shadowbanned for their political viewpoints and for election/health disinformation (see Figure 6). A 2019 survey from Pew Research Center found that 90% of Republicans believe it is likely that social media sites censor political viewpoints, as opposed to 59% of Democrats (Vogels et al., 2020).

GENDER IDENTITY	Social Media User Demographics (n=1006)	Reportedly Shadowbanned Social Media User Demographics (n=274)
Man	44%	54% ⬆
Woman	53%	41% ⬇
Non-Cis/ Non-Binary	2%	5% ⬆
AGE		
18-24	14%	19% ⬆
25-34	25%	23%
35-44	21%	26% ⬆
45-54	15%	19%
55-70	25%	13% ⬇

POLITICAL VIEWS	Social Media User Demographics (n=1006)	Reportedly Shadowbanned Social Media User Demographics (n=274)
Democrat	32%	26% ⬇
Independent	31%	33%
Republican	23%	29% ⬆
RACE OR ETHNICITY		
Caucasian/ White	75%	80%
Hispanic	16%	27% ⬆
African American/ Black	13%	11%
Asian	2%	0% ⬇
Other	10%	9%

▲ **Figure 4. Comparison of demographics of social media users in general and those users who reported being shadowbanned.** Arrows (⬆⬇) indicate statistically significant differences between groups. Source - CDT National Survey of Social Media Users 2021

Other work has also shown that trans and queer social media users believe that their content gets disproportionately moderated as well (Salty, 2021; Van Horne, 2020). Haimson et al. found that transgender users report having their content removed for being "adult despite following site guidelines, critical of a dominant group (e.g., men, white people), or specifcally related to transgender or queer issues" (2021, p. 466). In our own survey we found that trans and non-binary social media users were twice as likely as cis social media users to believe they had been shadowbanned.

Haimson et al. also found that Black social media users reported higher rates of content removal, particularly over issues related to racial justice or racism (2021). Our survey did not find that Black social media users reported being shadowbanned at higher rates, but they did report receiving harsher moderation actions than other users. Black users who were shadowbanned far more often had their entire accounts blocked (33% of Black shadowbanned users, versus 13% and 16% of white and Hispanic users respectively), as opposed to having specific posts or comments blocked.

"It seems like a strange time for me to get shadowbanned. My posts have felt pretty gentle lately. The only thing that could be controversial lately is my post for Fred Hampton's birthday, and I don't think it's that controversial." (Shadowbanned social media user, Interview, 2021)

However, unlike Haimson et al. (2021), our survey found that Hispanic social media users were significantly more likely to believe they had been shadowbanned. Hispanic social media users also had the widest gender gap in reported shadowbanning — 61% of Hispanic respondents who reported being shadowbanned were men compared to 35% women. We had not anticipated these results based on previous research, but conversations we had with experts on Latinx content moderation pointed to some possible explanations. Automated social media content analysis tools often train only on English language data (Elliott, 2021), and mixed language data often gets thrown out of training sets (Duarte et al., 2017, p. 15), so messages that mix Spanish and English may accidentally get flagged for shadowbanning more often. Another hypothesis is that Latinx culture and expression can use particularly melodramatic language (Flores-Saviaga & Savage, 2018; Sadlier, 2009), so algorithms trained without that cultural context may flag culturally acceptable behavior as inappropriate.

Finally, there are many smaller communities that we could not capture in a demographically representative poll that our research and interviews suggest may experience disproportionate shadowbanning. Sex workers, for example, have done extensive autoethnographic research on their own shadowbanning experiences (Fitzgerald & Sage, 2019; Valens, 2020). Hacking//Hustling, a sex worker advocacy group for sex workers' rights online, found that 31% of sex workers, 13% of activists, and 51% of sex worker activists report having been shadowbanned (Blunt et al., 2020). Are's autoethnography on pole dancing (which is distinct from sex work) documents how pole dancers similarly face disproportionate shadowbanning in what she calls a "shadowban cycle": whereby service providers attempt to demonstrate to the public that they take content moderation seriously by targeting the highly visible and easier to target women's bodies instead of the more difficult problems of hate speech (2021). Many groups besides sex workers and pole dancers also complain of disproportionate shadowbanning, including plus sized people showing skin (Joseph, 2019; Richman, 2019), midwives and other birth workers (Akpan, 2020), and international human rights organizers, especially in India (Ravi, 2021) and Palestine (7amleh, 2021; Human Rights Watch, 2021).

Which social media services shadowban and why?

Our survey also found wide variety in people's perceptions of being shadowbanned across different social media services. The results fall into three distinct buckets. On the high end of perceived shadowbanning is Facebook, where 8.1% of respondents who used Facebook believed they had been shadowbanned. In the middle are Twitter, Instagram, and TikTok, where 4.1, 3.8, and 3.2 percent of users respectively believed they had been shadowbanned. For all other social media companies (e.g. YouTube, Discord, Reddit, Pinterest), around one percent or less of users of those services perceived that they had been shadowbanned. It is important to note that this is a survey of users of different platforms, not necessarily posters of content. TikTok and Twitter, for example, may have a lower percentage of users that post content than Facebook, so the percent of users who could be shadowbanned may also be lower.

▼ **Figure 5. Perceived Shadowbanning Among Users of Each Social Media Service.** Source - CDT National Survey of Social Media Users 2021

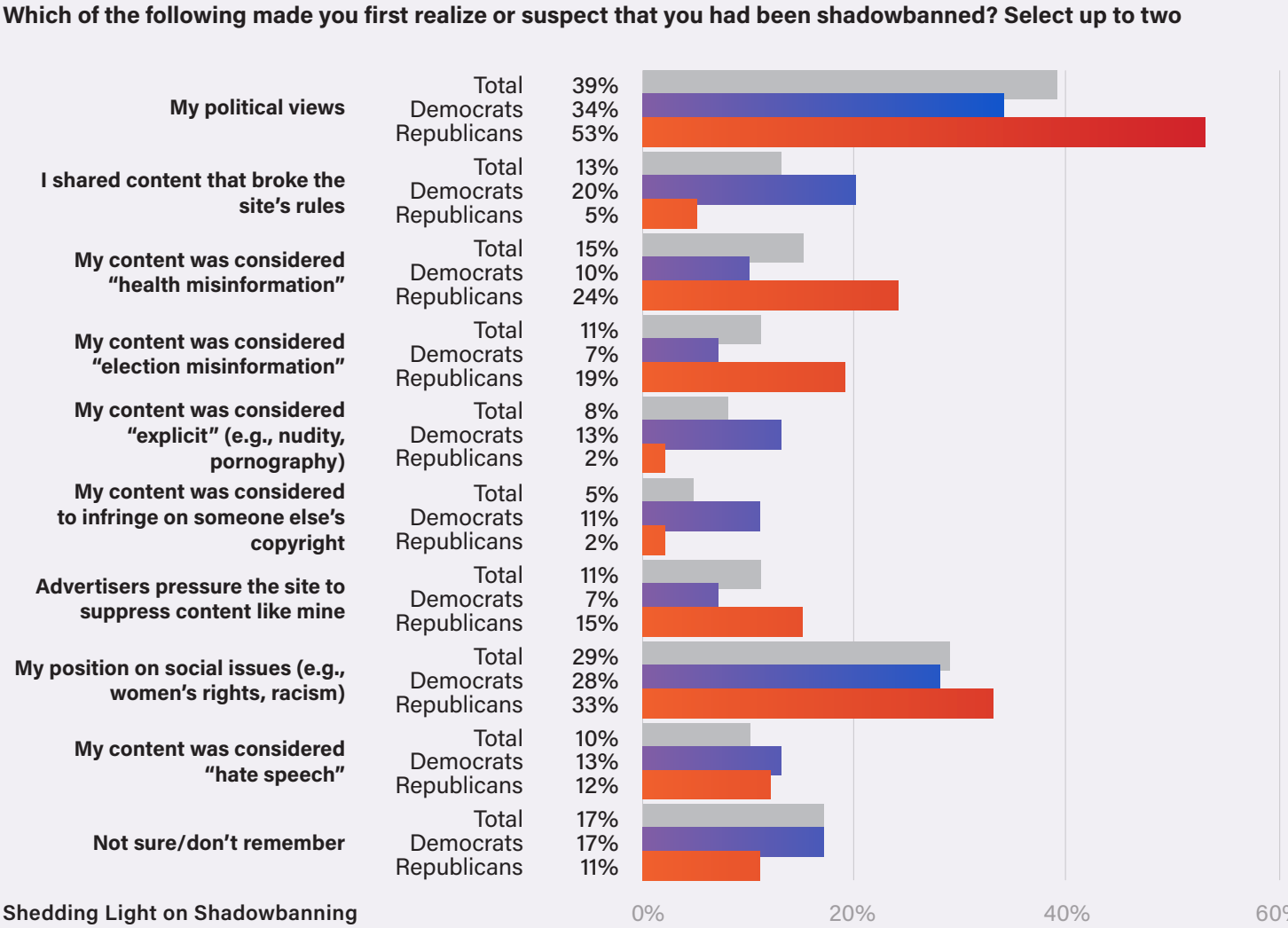
Social Media Service	Social Media Service Usage (n=1006)
YouTube	95.1%
Facebook	90.9%
Instagram	65.8%
Pinterest	57.6%
TikTok	49.4%
Twitter	48.2%
Snapchat	45.7%
LinkedIn	40.6%
Reddit	39.0%
Discord	22.6%
Twitch	22.5%
NextDoor	21.0%
Tumblr	19.1%

Social Media Service	Perceived Shadowbanned among Users of Each Service (n=274)
Facebook	8.1%
Twitter	4.1%
Instagram	3.8%
TikTok	3.2%
Discord	1.3%
Tumblr	1.0%
YouTube	0.9%
Twitch	0.9%
Reddit	0.5%
NextDoor	0.5%
Pinterest	0.2%
Snapchat	0.2%
LinkedIn	0.2%

Users also had a range of beliefs about the reasons the platform may have had for shadowbanning them, and those beliefs differed between Democrats and Republicans. Political views and position on social issues were the two most common reasons (39% and 29%) users reported having led to their being shadowbanned, though health misinformation, election disinformation, hate speech, and explicit content were also relatively common (15%, 11%, 10%, 8%). Republicans significantly more often believed they were shadowbanned for their political views, health misinformation, and election misinformation. Democrats significantly more often believed they were shadowbanned for explicit content and breaking the service's rules.

TikTok (10%) and Snapchat (11%) users who reported being shadowbanned were more likely to say they believed that it was because their content was explicit (e.g., it contained nudity) compared to users of other platforms. Redditors (22%) were more likely to suggest that their perceived shadowbanning experience was because of pressure from advertisers on their platform to suppress certain content. Meanwhile, 17% of users were not sure or did not remember why they were shadowbanned.

▼ **Figure 6. Reasons social media users gave for why they believed they had been shadowbanned, (% of those who say they were shadowbanned, n=274).** Source - CDT National Survey of Social Media Users 2021



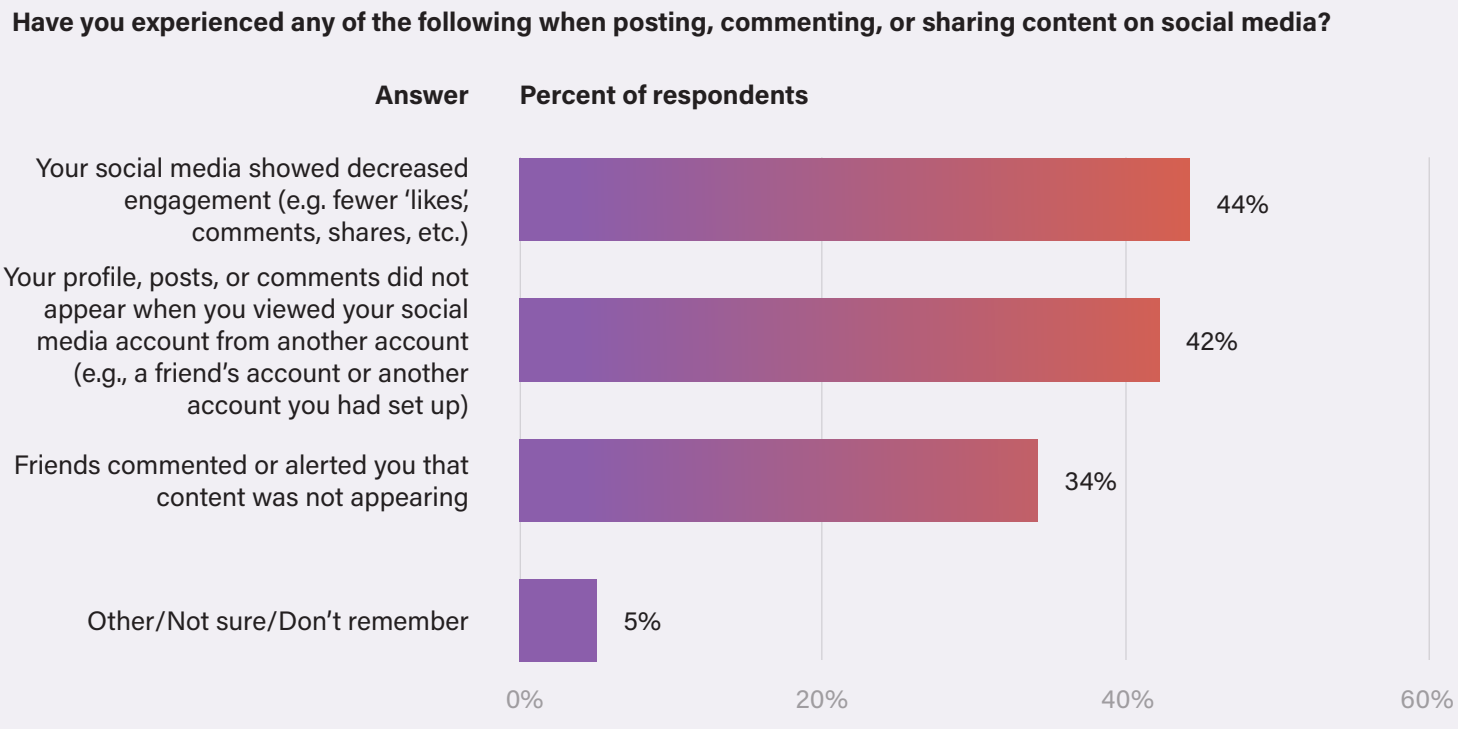
How do users diagnose and respond to their own shadowbanning?

How can users tell when they have been shadowbanned? They have no way of knowing for sure — such is the opaque nature of shadowbanning — but there are a few common tactics available to them to make an educated guess.

Our survey found that the most common method (44%) that users use to diagnose their own shadowbanning is to look for a drop in their engagement metrics. Several people we interviewed also mentioned that they saw a precipitous drop in their engagement — sometimes as much as a ten- or even one hundred-fold decrease between one day and the next. Some fluctuation in engagement is to be expected, and social media company workers interviewed pointed out that users who post similar content over time may naturally get less engagement as people move onto something new. Still, sharp dips raise questions. For example, the Twitter user @s8n, a Satan parody account that for months was growing by thousands of followers a day, went from adding almost 6,000 new followers one day to only 217 just five days later (Oremus, 2019a). Twitter eventually attributed the problem to a bug (Oremus, 2019b).

Our survey found that the two other most common methods by which users diagnosed their own shadowbanning were to use another social media account to test whether their content was hidden (42%) or learn from friends that their content was not appearing (34%). This process can be difficult with feeds and recommendation algorithms, since there is no objective baseline of what should appear, but missing comments and search results can be easier to check using these methods.

▼ **Figure 7. What made social media users say they were shadowbanned (% of those who say they were shadowbanned, n=274).** Source - CDT National Survey of Social Media Users 2021



One strategy that we didn't ask about in the survey was use of shadowban detection tools. Programmers have created tools that automate the process of checking how posts appear from external accounts. Examples include r/CommentRemovalChecker for Reddit (MarkdownShadowBot, 2018), the now-defunct Triberr for Instagram (Triberr, n.d.), and shadowban.eu (shadowban.eu, n.d.), which is defunct as of January 2022. These tools are likely very popular — one shadowban tool developer told us theirs was used 130 million times before it shut down. r/CommentRemovalChecker has been used at least one hundred thousand times, and likely well more than that (r/commentremovalchecker, n.d.).[3] Importantly, these tools can only catch forms of shadowbanning that can be determined in an automated, disprovable way, so they could not, for example, tell if a user's content has been algorithmically suppressed.

Nearly half of shadowbanned respondents (42%) could not find any means of recourse for having been shadowbanned. However some were able to find a way to report a problem with their account (14%), or found a form (8%) or email (7%) through which to reach out. Of those who were given options to resolve or found means of recourse (n=145), 65% (n=94) attempted to use them and 73% of those who used them resolved the issue. But many simply waited for the shadowban to be lifted. One study on Facebook Pages, for example, found that shadowbans came in units of seven days (Horten, 2021). Facebook Page admins would report that they would see 93-99% decreases in engagement metrics that would last one, two, or as many as eight weeks.

"I know this sounds kind of tin-foil hatty but . . . when you get a post taken down or a story, you can set a timer on your phone for two weeks to the godd*mn f*cking minute and when that timer goes off you'll see an immediate change in your engagement. They put you back on the Explore page and you start getting followers."
Shadowbanned Instagram user (**Constine, 2019**)

3 r/CommentRemovalChecker works by having users post a link to a comment they suspect is shadowbanned and a bot checks to see whether or not it is visible from an external account. This number was calculated by using the Reddit API to count how many posts exist on the r/CommentRemovalChecker subreddit. However, the bot allows users to automatically delete their requests after it determines an answer, so this number may be undercounting.

What mitigation tactics do users employ?

Just as it is difficult for users to determine whether they've been shadowbanned, it is difficult to figure out what to do in response and how to prevent it in the future. Without authoritative explanations from social media companies, individual communities develop their own folk theories about why and how shadowbanning and other forms of content moderation occur (Eslami et al., 2015; Karizat et al., 2021). Folk theories can have utility even if the underlying reasoning is flawed (Kempton, 1986), and, throughout our interviews, we found that users developed their own internally consistent communities of practice. As one Black woman who was shadowbanned put it:

> What I have known — and this might be my circle — but the people I've seen this [shadowbanning] happen to and the people who have offered the most valuable support are 100% Black women. When I'm sharing the error messages that I'm getting, I can see the Black women saying "you gotta do this, you gotta do that" or "this happened to me last week"...Other people are trying to send me tech articles say, "Trying clearing your cache." And I say, "Oh sugar..."
> *Source - Interview with a shadowbanned social media user - September, 2021.*

Below, we list a few tactics that interviewees mentioned to avoid shadowbanning. In order to protect the utility of these methods, we avoid naming specific tactics that interviewees employed and will only give examples that have been published publicly.

STEGANOGRAPHY AND MISSPELLINGS

Steganography is the practice of hiding secret messages inside of something that is not secret. Users interviewed reported using steganography and misspelling certain words to avoid having their content flagged by content analysis algorithms. While this is a response to all forms of content moderation and not just shadowbanning, if users suspect that a social media service shadowbans, they may hide more words than they otherwise would because they believe that they will be unable to verify which words do and do not trigger automated moderation. People and businesses that post about cannabis for example will often replace letters, using words like "c*nnabis," "w33d," and "st0ner" (Bartlett, 2021). Other communities such as the pro-eating disorder and non-suicide self injury communities, have gone further off the beaten path, using less intuitive hashtags that only in-group users understand (Gerrard, 2018). As one interviewee put it, "Sex workers in particular will type l33t speak like it's 2002." Some users use far more involved methods to avoid shadowbanning though, such as Feroza Aziz, who disguised a criticism of China's treatment of the Uyghur people as a makeup tutorial (2019).

DISTINGUISHING FROM INAUTHENTIC BEHAVIOR

"There are all these strategies that are in the sex work zeitgeist — make sure you sprinkle in a bunch of every day stuff with your advertising so not all of your tweets look like advertising, make sure you don't look like bots. But that's surprisingly hard to do so we've come up with all these ways to advertise without looking like bots. It's a job in itself just to navigate these systems" (Shadowbanned social media user, Interview, 2021)

Multiple shadowbanned users we spoke to said that they intentionally tried to distinguish themselves from bots. Users reported disconnecting third-party apps that could automatically post on their behalf and intermingling original and reshared content. Other users reported removing irrelevant and extraneous hashtags, which they were previously using to increase the discoverability of their content. Removing extraneous hashtags is the most common advice that bloggers give to users looking to avoid being shadowbanned (Forsey, 2021; McLachlan, 2021; Zhang, 2018), but some social media services have explicitly said the number of hashtags a user uses is not considered when determining content moderation actions (Constine, 2019).

APPEASING THE ALGORITHM

Interviewees reported posting content that they believed would be rewarded by recommendation algorithms and removing potentially objectionable content they believed it would punish (Cotter, 2019). Sex workers, for example, mentioned removing links to OnlyFans in their bios and posts to avoid being shadowbanned. Multiple conservatives we interviewed mentioned that they routinely deleted their posts so that if someone they tagged in the past became a persona non grata on the social media service, they would not have their own content shadowbanned by association.

"The algorithm prefers I post a sweaty selfie." (Shadowbanned social media user, Interview, 2021)

Interviewees tended to have a strong sense of what social media recommendation algorithms "liked," and chief among them was images of faces. As one animal rights activist interviewed put it, "I've literally had people say, you aren't using the algorithm properly. I know Instagram loves faces, cute animal photos, but that's not what I'm using it for. I'm using it to educate people." Organizers and advocates we talked to mentioned the strategy of posting political messages alongside photos of their faces. Research on Facebook Pages similarly found that certain Pages could be affected by bans where posts containing links would not show up on other users' feeds while posts consisting of images would (Horten, 2021).

SWITCHING TO MORE PROTECTED FORMS OF SPEECH

To avoid shadowbans and other content moderation reprisals, some users interviewed switched to more protected forms of speech. In particular, they shifted from advocating about controversial topics online to educating about them. Sex workers interviewed discussed how other community members have shifted to sex education. Similarly, cannabis advocates will often frame their work as education, such as the Pot Brothers At Law, a TikTok account, which informs users about their rights regarding cannabis (Pot_Brothers_at_Law, n.d.). Similarly, after Facebook cracked down on anti-vaccine groups after the 2014 Disneyland Measles outbreak, anti-vax rhetoric shifted from telling others not to get vaccinated to promoting users' civil liberties and their right to choose whether they get vaccinated (Broniatowski et al., 2020). Conservatives we interviewed mentioned using similar strategies to be able to talk about Covid-19 without getting shadowbanned.

What are the effects of shadowbanning?

The practice of shadowbanning has some utility for improving content moderation outcomes, but like other forms of moderation, it can also do harm to those directly affected, and to the service's users more broadly. The marginalization that stems from shadowbanning can be difficult to see or interrogate. However, our survey and interviews with shadowbanned social media users revealed a range of potential harms for individuals, groups, and society as a whole.

What harms does shadowbanning do to individuals?

SADNESS, ISOLATION, AND EMOTIONAL HARM

Much of modern socialization occurs on social media, and users who cannot be heard or seen by others online often experience feelings of sadness and isolation. In our survey, 54% of shadowbanned users said that being shadowbanned made them feel isolated and removed from their social group, community, or society at large. Even more (65%) said that shadowbanning made them less able to connect with new social groups or communities of interest. One interviewee, for example, had a friend who received a nomination for a prestigious award and she felt hurt that her shadowban prevented her from congratulating the nominee.

"First I was confused, then I was frustrated, then there was even a little bit of shame. Is there something that I did? Did I do something wrong? Do they think I'm a spam bot? Do they think I bought followers? There's something that makes me feel marked." (Shadowbanned social media user, Interview, 2021)

Other forms of content moderation have harms. Myers West, for example, found that users who have their content removed or their accounts banned often feel isolated because they cannot share personal news or feel completely cut off from services with few alternatives, such as Spotify and Tinder (2018). However, shadowbanned users, unlike explicitly banned users, cannot share a screenshot to another social network of the service taking action against them. And in ranking and recommendation systems, shadowbanned content is indistinguishable from unpopular content.

"People whose content just sucks will also say they're shadowbanned. So it's hard to get people to believe it's actually happening." (Shadowbanned social media user, Interview, 2021)

Shadowbanning can also be a uniquely isolating experience because many users either do not know about it or do not believe that it occurs. Nearly half of social media users in our survey had neither heard of the term nor knew shadowbanning occurred. And Cotter found that some

of those who are familiar with the term doubt it is real. As one person interviewed for that study said, "I found everybody was up in arms about shadowbanning, because if you use too many hashtags, or the same hashtags, you'll get shadowbanned, blah, blah, blah. And I was like, 'No, that's not right.' Instagram never publicly announced that they were shadowbanning. It was just all hearsay, so it was completely BS" (2021, p. 12).

The isolation shadowbanned users feel could be exacerbated in the more immersive experience of virtual reality. Operators of virtual realities will face strong incentives to moderate strictly: as Meta CTO Andrew Bosworth said in an internal memo, virtual spaces may have a "stronger bias towards enforcement" to prevent abuse (Robertson, 2021). VR companies will also face competing incentives to keep new users engaged and to demonstrate growth in these environments, which may lead them to consider shadowbanning as a useful tool. Sony already has an extensive patent on "Shadow banning in social VR setting" that lists many ways it could be implemented (Miyaki, 2021). An early build of Meta's Horizon Worlds, a VR social space, had a version of shadowbanning implemented.

"Both the blocker and the blocked were made invisible to one another, but allowed to continue interacting with the same virtual world. While they couldn't see one another, they could see each other's effects on their shared environment. If someone blocked you, your obscene gestures might be invisible to them, but you could still move the furniture about and rattle chains — practically becoming a poltergeist." (**Duffield, 2021**).

FINANCIAL HARM

Shadowbanning can cause financial harm to users who depend on social media for their income. In our survey, 20% of shadowbanned users indicated that being shadowbanned affected their ability to make a living. In our interviews, we found that this affects educators, artists, pole dancers, and sex workers in particular. Are (2021) wrote specifically about how shadowbanning inhibits pole dancers from using social media to teach and promote fitness classes and performances. Shadowbanning contributes to what Duffy calls "algorithmic precarity, which is "the turbulence and flux that emerge as a routine feature of platformized labor" (2020, p. 2). Are sees this largely as a way for platforms to externalize costs — "Institutions and businesses ineffectively attempt to reduce risks for their citizens or customers by restricting civil liberties. This way, corporations attempt to avoid undesirable effects by arbitrarily identifying risks to prevent, increasing the marginalisation of society's 'others'" (Are, 2021, p. 2).

MORE EFFECTIVE TROLLING

Shadowbanning is designed as an anti-troll measure, yet trolls have weaponized for their own ends the very automated methods of content moderation designed to stop them. Trolls will often engage in bad faith attacks on vulnerable populations (e.g. Colombo, 2021) by bombarding targeted users with false reports of content policy violations, eventually triggering automated content moderation actions against the targeted user

(Clark-Flory, 2019). If a user facing such abusive reports is informed that their content has been moderated, they can appeal to the service operator with evidence of the attack. Shadowbanned users, however, are not given a content moderation decision to dispute, and may have nowhere to bring evidence of an organized attack.

In some cases, the very act of shadowbanning can serve as a rallying point for trolls. For example, in 2013, users from the not-yet-banned Reddit community r/n***ers repeatedly attacked the r/blackgirls community with racist comments and soon found themselves shadowbanned (Todd, 2013). Users would watch their "comrades" disappear from conversations and celebrate their sacrificial shadowbans with song posts in a separate community made just for that purpose, r/RedditMartyrs (Reddit, n.d.).[4]

4 As of this writing, r/RedditMartyrs is still up, though it has been inactive since 2013.

What harms does shadowbanning have on groups?

In reality, FOSTA/SESTA raise clear threats to the constitutionally protected speech of sex workers and others. For more of CDT's opinion on FOSTA/SESTA, see (**Woolery, 2018**).

EXCLUDES VOICES FROM CONVERSATION

Communities that found themselves repeatedly shadowbanned also found themselves systematically excluded from larger conversations happening on social media. One sex worker interviewed gave the example of a tech conference they went to where attendees were discussing FOSTA/SESTA, legislation then under consideration by Congress (and now enacted into law) that was purportedly designed to limit sex trafficking online. Conference attendees were discussing the bill in a shared Twitter hashtag, and, by and large, people were in favor of the bill. The interviewee tried to voice their disagreement by using the hashtag, but when checking their posts elsewhere, they found out that their posts and another sex worker attendee's posts were not appearing.

In our survey, a majority of all social media users (79% of users who reported having been shadowbanned and 57% of those who had not reported being shadowbanned) agreed that with shadowbanning, not all perspectives and points of view are adequately represented on social media. Shadowbanned users we interviewed expressed similar frustration. Conservative interviewees in particular expressed frustration at not being able to respond to current events. As discussed earlier, shadowbanning is often done to borderline content. When social networks are not sure if a piece of content breaks their rules, they frequently reduce its distribution or take some other undisclosed action short of removal. As one interviewee complained, by the time they found out that they were shadowbanned, complained about it to a service provider, and had the issue resolved, the shadowbanned post they had made was no longer timely and therefore unlikely to draw attention.

UNABLE TO FIND COMMUNITIES

When a hashtag gets shadowbanned, relevant communities may not be able to find one another. As one interviewee said, "If I'm in Kentucky and I'm trying to come out, I can't find [the LGBTQ community] on social media, because #gay, #bi, #lesbian are being tagged as inappropriate." Sex workers and activists also mentioned how shadowbanning made it more difficult for them to find and share strategies for staying safe (Blunt et al., 2020). In interviews, social media users from communities that have reappropriated slurs or other harmful language also expressed difficulty with forming community bonds. The unique effect of shadowbanning is that users who do not understand it as a moderation technique or know that it is happening may feel isolated and come to believe that such resources do not exist, rather than the fact that the social media service refuses to host them.

SHADOWBANNING BY ASSOCIATION

According to our survey, 74% of social media users believe that shadowbanning unfairly targets certain groups of people more than others. This belief is impossible to validate: while there is research indicating that some groups report experiencing more content moderation than others (Blunt et al., 2020; Haimson et al., 2021), independent researchers cannot tell if that reporting is true. And even if it were, researchers could not determine whether these groups actually violate community guidelines more than others, whether those community guidelines are biased against certain groups, or whether it is appropriate for guidelines to be biased against certain groups if it means mitigating harms.

Still, research suggests that at least on some social networks, shadowbanning frequently occurs in clusters of associated users. A study from Le Merrer et al. ran a network analysis using a Twitter shadowban detection tool and found that users who had interacted with someone who had been shadowbanned were nearly four times more likely to be shadowbanned themselves (their chances going up from 2.3% to 9.3%) (2021). These shadowbans, they argued, didn't follow the pattern of a random software bug, to which a Twitter spokesperson had attributed the bans (Gadde & Beykpour, 2018), but rather followed the pattern of a disease that could "spread" between users that interact with one another.

Though it is hard to disentangle cause from effect, there is some evidence that Twitter's automated content moderation tactics are at least contributing to this phenomenon. In a 2018 post-mortem after its debacle with allegedly shadowbanning Republican politicians, Twitter explained that the system's inputs included, "who you follow, who you retweet...who follows you, who retweets you" (Gadde & Beykpour, 2018). Though that system was modified, Twitter again admitted that it opaquely moderates accounts that associate with supposedly low quality, "spammy" accounts, when commenting on how it accidentally shadowbanned a Satan parody account. As one journalist described it, "Accounts that hadn't actually broken Twitter's rules themselves could still find themselves filtered from public view through a kind of guilt by association" (Oremus, 2019b).

Content moderation by association can be an effective tactic for detecting harmful content such as coordinated information campaigns or fake engagement (Pacheco et al., 2021; Serrato, 2020), so other social media services also likely use such tactics but are not as open about it. Services may opt for moderation tactics that they do not communicate back to the user, including reducing the distribution of content, since it is more difficult for advanced actors to engage in countermeasures against these tactics.

Meta, for example, often reduces the distribution of content flagged for inauthentic sharing, pages predicted to be spam, and suspected disinformation outside of health, elections, and manipulated media (Meta Transparency Center, 2022). However, moderation-by-association may accidentally sweep up whole groups of authentic users, particularly those who may act in ways that can be confused as spam (e.g. by sending high volumes of identical messages), such as activists (Starbird et al., 2019). Moderation-by-association could cause any bias that exists in content moderation algorithms to build upon itself.

What harms does shadowbanning have on society?

"You have to separate fact from fiction a little bit. You have people saying, 'My tweet only has five likes, I'm being shadowbanned.'" (Shadowbanned social media user, Interview, 2021)

DISTRUST AND CONSPIRACY THEORIES

Conspiracy theories thrive in secrecy (Moynihan, 1998; Pozen, 2009), and shadowbanning is secretive on two levels — users are kept in the dark about when their content is moderated and social media services do not disclose that they practice shadowbanning. The combination of these two forms of secrecy has contributed to a large, deep public distrust of social media services and their content moderation practices. Many users who believe they have been shadowbanned have likely not been shadowbanned, and instead simply have unpopular content. However, the opaque nature of shadowbanning often makes it impossible to distinguish between unpopular content and shadowbanned content, and when users have the chance to blame the service, they will likely take it. As another interviewee described shadowbanning, "It has metastasized into an unfalsifiable catch all."

Social media thereby often gets cast as a shadowy cabal and its content moderation practices described as a means to push a certain agenda. As U.S. Rep. Jim Jordan (R-OH) put it, "Big tech is out to get conservatives. That's not a suspicion, that's not a hunch, that's a fact" (Merlan, 2020). In interviews, social media users from many groups expressed this sentiment, but conservative social media users did so in particular. For example, multiple conservatives interviewed believed that social media services used "bugs" in their algorithms as an excuse to hide conservative content. As one interviewee put it, "You can always use the classic Google response — oh it's just a technical problem. [Laughs.] It's always a technical problem, and almost always the technical problem hurts conservatives."

A canonical example of this is Twitter's quality filter debacle, discussed above, in which a Vice article found that many people, including prominent Republicans, had their usernames hidden from Twitter's autofill in search (Thompson, 2018). The controversy resurfaced a selectively edited video that Project Veritas had released a few months earlier that falsely made Twitter employees appear like they were admitting to anti-conservative bias in their content moderation practices (Lee, 2018). Even though Twitter presented evidence that the quality filter problem was actually a bug, (Gadde & Beykpour, 2018), the Vice article and the Project Veritas piece had already whipped up political controversy. Former U.S. President Donald Trump captured Republican sentiment when he tweeted, 'Twitter "SHADOW BANNING" prominent Republicans. Not good. We will look into this discriminatory and illegal practice at once! Many complaints.' (Samuels, 2018).

"BLACK BOX GASLIGHTING"

"Black box gaslighting" (Blunt et al., 2020) is when social media services use their position as the sole authority of their own algorithms "to undermine users' confidence in what they know about algorithms and destabilize criticism" (Cotter, 2021, p. 5). In the case of shadowbanning, companies appear to deny or mislead users about how and whether they moderate their content without users' knowledge and shift blame to the individual user. In 2017, for example, in response to users complaining that certain posts were not appearing in hashtag search, Instagram wrote, "Having a growth strategy that targets the right audience is essential to success on Instagram. Good content on Instagram is simply good creative" (Instagram, 2017) (Instagram has since admitted it hides certain sensitive content within hashtags (Constine, 2019)).

"Shadow banning does not exist, it is a persistent myth …I personally think it sounds kind of cool and sexy so people love saying it but the day that the shadow banning word became a thing, it's because there was legitimately a bug that was affecting hashtags" - Director of Fashion Partnerships at Instagram, Eva Chen (**May, 2019**).

Social media services similarly mislead users when they deny that they shadowban by implicitly or explicitly using a definition of the term "shadowban" that is far more narrow than the way people use the term today. In their post, "Setting the record straight on shadow banning," Twitter, for example, adopted the old, classical definition of shadowbanning to be "deliberately making someone's content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster" (Gadde & Beykpour, 2018). And while Instagram acknowledged that shadowbanning is "a broad term that people use to describe many different experiences they have on Instagram," instead of saying which of those experiences they do and do not engage with on their service, they shifted blame for content not appearing in feeds and recommendations back to the end user, saying, "We get that people get confused why their content isn't popular" (Mosseri, 2021a). Users we interviewed described this experience as akin to gaslighting.

"It feels like gaslighting. I'm loath to use the term, cause it doesn't feel like anything, but it's gaslighting." (Shadowbanned social media user, Interview, 2021)

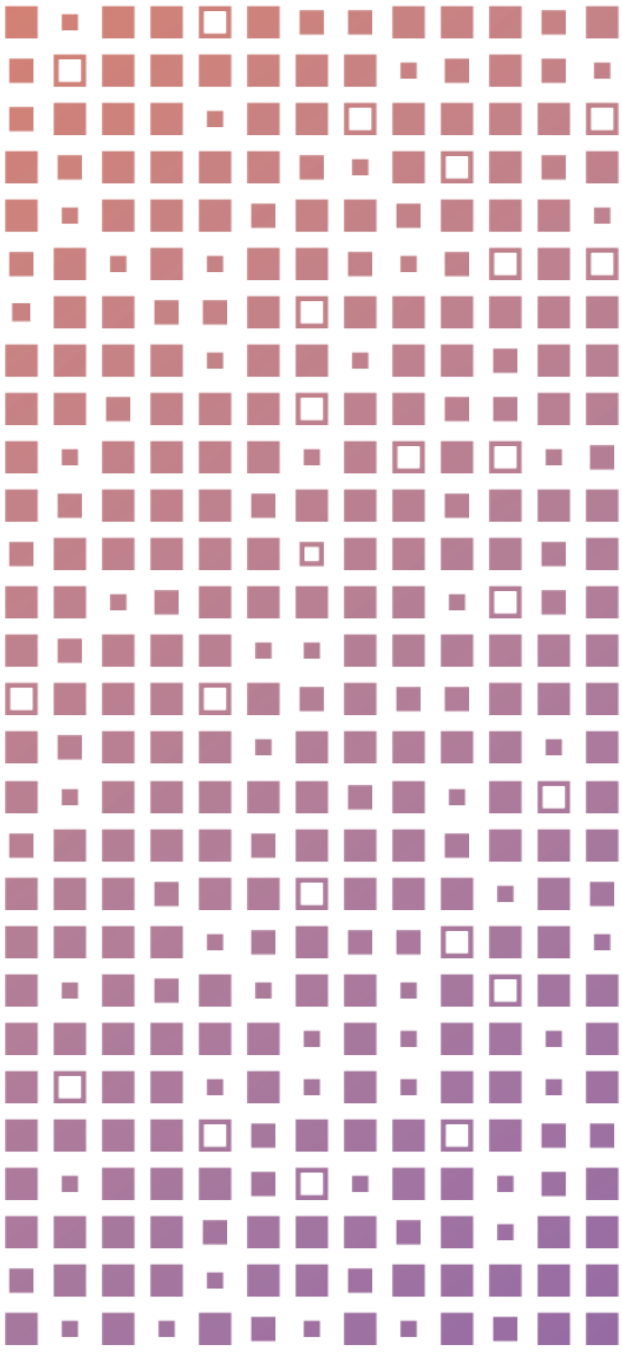
INABILITY TO CORRECT CONTENT MODERATION ERRORS

Whether because of bias, the difficulty of content moderation at scale, or any other reason, service providers inevitably moderate content that they shouldn't or don't intend to (Sheppard, 2021). To mitigate this, service providers rely on — and publicize — their mechanisms for appealing moderation decisions. But these mechanisms cannot correct misguided shadowbanning since users are not informed of a decision to dispute. Sometimes, media or social media attention for a controversial content moderation action against a popular social media account can also act as a corrective. For example, when plus-size model Nyome Nicholas-Williams repeatedly had her photos deleted and her account blocked on Instagram, the hashtag #IWantToSeeNyome became popular

across multiple social media services, and news outlets discussed how Instagram's enforcement of its anti-nudity policies may be biased against Black and plus-size users (Fleming, 2021). In response, Instagram changed its policy on breast holding (Davis, 2020). Nicholas-Williams was only able to know for sure that her content was being moderated because she was informed that her content was deleted and her account was blocked. Had Instagram instead shadowbanned her account, and had she not been able to prove that her content was being moderated, it likely would not have gotten the swell of public attention that led to a change of Instagram's policy.

Social media services are especially likely to make mistakes when shadowbanning borderline suspected content. Borderline content is challenging to detect because most borderline content policies are confusingly worded and vague in scope. Decisions to shadowban borderline content may therefore be more likely to be erroneous than other content moderation decisions. All in all, systematic errors can more easily proliferate without being noticed by researchers, journalists, or affected communities themselves since users, communities, and others have little way to know it's happening nor an easy way to prove it to others.

Recommendations



In this paper, we reviewed what shadowbanning is, who it affects, and how it can harm users. We now conclude with recommendations for social media companies on how they can mitigate some of these harms with changes in policies and practices. We urge online service providers to shed light on shadowbanning by doing three things:

1. Disclose to users and the public whether they use opaque content moderation practices and in what circumstances;
2. Minimize the circumstances in which they use shadowbanning or opaque content moderation practices; and
3. Conduct and enable further research into the effects of opaque content moderation techniques, including through transparency reporting and providing independent researchers with access to moderation data.

Publish shadowbanning policies

Some of the harmful effects of shadowbanning don't come from the actual undisclosed removal or reduction in distribution of content, but rather from the fact that social media companies fail to acknowledge in general, or even outright deny, that they reduce or remove content without informing end users at all. This meta-secrecy can lead to end users feeling "gaslit" (Cotter, 2021) and likely contributes to the isolation shadowbanned users describe feeling in our survey. It also likely fuels larger societal distrust of social media companies and conspiracy theories about how they engage in content moderation.

Online services should publicly disclose their policies around opaque content moderation practices (regardless of whether they use the term "shadowban"), including whether they ever take action against users' content or accounts without informing them. Providers should explain the general criteria they apply for deciding when to engage in opaque content moderation. Producing such a disclosure will require online service providers to ensure that they have clear and consistent internal policies around the practice, and will help clarify to users when shadowbanning is not occurring. Our research also found that many users already assume, possibly incorrectly, that widespread shadowbanning is affecting themselves and their communities. In addition to policy disclosures, social media companies should include in their transparency reports basic data about how many accounts and posts they moderate without disclosing to the user.

In circumstances where shadowbanning may be justified, such as stopping spammers or preventing trolls from creating sockpuppet accounts — general disclosures will not defeat the efficacy of the techniques. Shadowbanning is effective in these circumstances because it is difficult or, in the case of suppressed algorithmic distribution, nearly impossible for the affected user to confirm in their individual case. But an online service's general shadowbanning policies must be able to withstand public scrutiny; if an online service fears reputational risk from disclosing its policies about opaque content moderation, that is a signal it should revise or abandon those policies entirely.

The only user-centric justification we heard social media companies offer for secrecy about whether they ever engage in opaque content moderation practices is that authoritarian governments may use a company's acknowledgment that it engages in shadowbanning to demand that they shadowban particular content. This is a risk for any technical approach to content moderation that a service provider employs and is not unique to shadowbanning. For any type of government demand to restrict content, providers should require that such demands follow established domestic legal processes and should interpret such demands so as to minimize the negative effect on freedom of expression (Global Network Initiative, 2019). Providers should also review their existing procedures for responding to government demands that include a gag order (i.e. an order not to inform the affected user that their content has been restricted), which share significant similarities with shadowbans. Providers should conduct human rights impact assessments to examine whether general disclosure of the provider's ability to shadowban is genuinely likely to yield more lawful orders to restrict content that are accompanied by gag orders in a given jurisdiction.

Don’t shadowban reflexively

In this paper, we do not call for a full stop to shadowbanning. Shadowbanning can be warranted when it prevents bad actors from structurally misusing or abusing a service, and it may be particularly helpful in protecting users from spam, coordinated disinformation attacks, and sockpuppet accounts from harassers. But when used to minimize objectionable content, the harmful effects of shadowbanning — how it feeds public mistrust, reduces accountability for social media services, and leaves users feeling isolated and gaslit, to name a few — are too great. Online services should err much more on the side of informing users about actions taken against their content or accounts, and reserve opaque content moderation techniques only for situations where informing a user about an action taken against them would meaningfully and demonstrably harm other users.

In our interviews, representatives from social media companies repeatedly cited the difficulty of communicating content moderation actions besides removal (such as reducing algorithmic distribution), given the broad range of digital literacy among end users. But this problem is only exacerbated by the prominence of opaque and unintuitive recommendation algorithms. We believe that communicating moderation actions to users, even algorithmic ones, is a solvable design problem, and providers should prioritize designing and developing their recommendation algorithms with transparency and explainability in mind.

There’s an entire research community dedicated to solving this problem, centered around the ACM Conference on Fairness, Accountability, and Transparency (FAccT).

Social media services should also help users understand when they are not being shadowbanned. One way to do this is to be explicit about when technical glitches are occurring. Multiple interviewed users believed that companies use “technical glitches” as an excuse to shadowban certain individuals, so if companies routinely publicly announce when errors are occurring — as Facebook and Instagram have rolled out programs to do (Ahmed, 2021; Instagram, n.d.-d) — users may also come to better trust social media companies’ claims about content moderation.

Conduct and enable research on the effects of shadowbanning

This paper is limited in its capacity to describe shadowbanning and its harmful effects because of the lack of available research on the topic. Only service providers themselves can give users and researchers the necessary data and information to develop a better understanding. Social media companies should make data available to independent researchers about what content they are shadowbanning in order to help the public understand what adverse effects it may have or what systematic, potentially harmful errors may be occurring.[5] In general, social media services share little data with independent researchers about how they moderate content, creating what one researcher called a "memory hole," an Orwellian term used to describe the disappearance of inconvenient information. Shadowbanning practices may have unintentional consequences that researchers do not even predict, as we experienced with the unexpectedly high rate of Hispanics who believed they were shadowbanned in our own survey. Our survey suggests that shadowbanning may also have previously unstudied gendered and racial dynamics, and sharing more data could help researchers better understand those interactions.

In particular, social media services should make available to researchers data that could help them understand if shadowbanning-by-association is really happening, and if so, who might inadvertently be getting caught up in it. Overzealous shadowbanning-by-association could lead to the silencing of entire groups, even members who have never posted abusive content themselves.

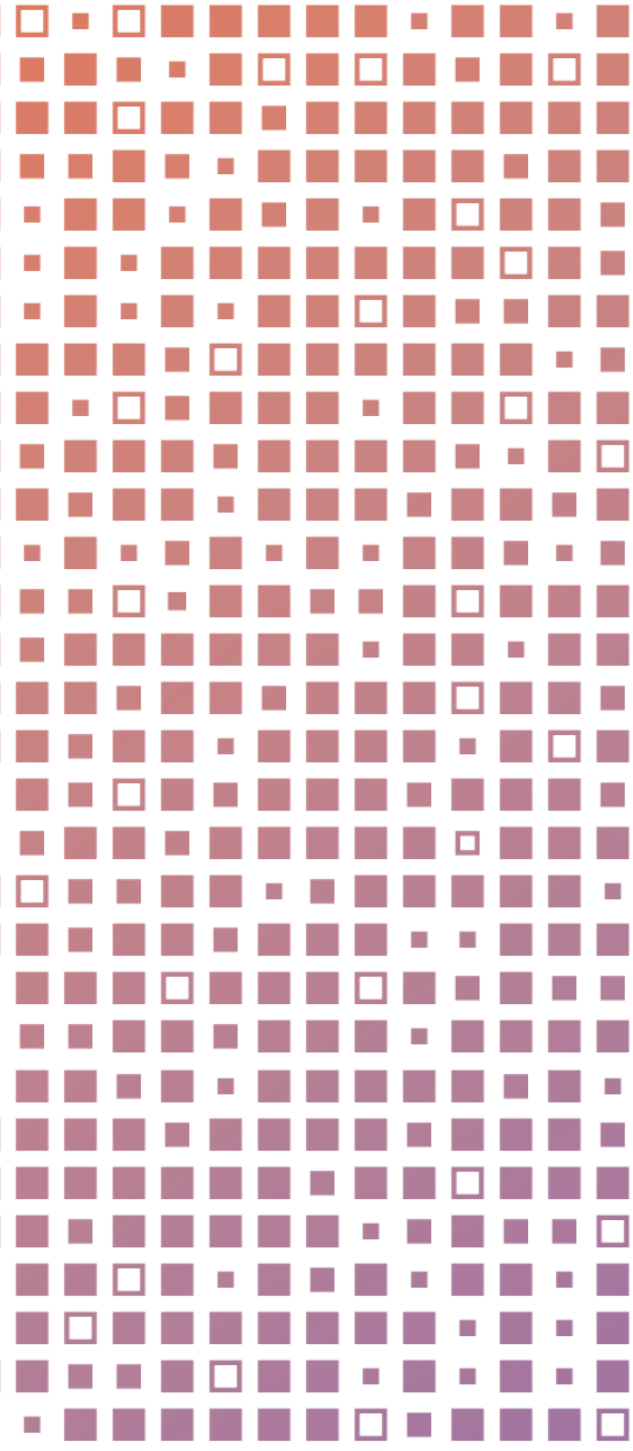
Each of these recommendations — to disclose shadowbanning policies, inform users about moderation decisions, and conduct and enable research into the effects of shadowbanning — will require online service providers to better communicate internally about their policies and practices that affect users' speech and access to information. Our research found that this is not always happening. In their corporate structures, social media companies often separate their Trust and Safety teams, who deal with content removal, from their algorithmic recommendations teams. In interviews with social media company employees, we found that content moderation actions that fall under the purview of Trust and Safety often get communicated back to the end user, while recommendation-based interventions do not, even when, from the user's perspective, their content's distribution is so diminished that it is as if it has been

5 Giving researchers access to data entails difficult tradeoffs between privacy and utility. See Vogus & Llansó's *Making Transparency Meaningful: A Framework for Policymakers* (2021) and CDT's forthcoming paper on lessons social media companies can learn about sharing data from other sectors for more.

removed (e.g. content appearing at the bottom of a feed or not in search results). As more services employ less-obvious content moderation actions, it will be vital for online services to have internal clarity about whether and how they restrict users' content, when they communicate those restrictions to users, and how they communicate those practices to the public.

Our goal here is to illuminate the meaning, practices, and effects of shadowbanning. However, it is also a call to action for online service providers, urging them to limit shadowbanning to a select few circumstances that they acknowledge publicly. Keeping users safe and keeping them informed about how their content is moderated should not be mutually exclusive, and the best way to improve both is by letting the light in.

Appendix A: Methodology



For this paper, we engaged in three forms of research: a literature review, semi-structured interviews, and a survey. For the interviews, we spoke to a total of 34 people between August 2021 and January 2022. Thirteen claimed to have experienced shadowbanning, thirteen worked at social media services, and eight were members of academia or civil society who worked on issues that relate to shadowbanning. The main topics that we explored in the interviews were individual and observed instances of perceived shadowbanning including types of content shadowbanned, impacts on the interviewee and community, means of recourse, and responses by the social media company. For each interview, we obtained informed consent and told interviewees that we would not reveal their names or employers, so as to protect their privacy and alleviate fears of reprisal. However, we did ask permission to inform interviewees that we may quote them and/or attribute quotes to them in more general terms (e.g. attribute quotes to "a Black activist" or "an employee at a social media company").

Finally, we commissioned an online, nationally representative survey of social media users in the U.S. to find out how many believed they had experienced what we call shadowbanning and what their experiences were like. The survey was administered in English between November and December 2021, by Edge Research. Our sample consisted of 1205 people with the results weighted based on age, race, and gender. We again obtained informed consent and have published the survey instrument and raw results data along with this report. It is important to note that with this survey, we are only able to probe how many social media users perceive themselves as having been shadowbanned, not how many have actually been shadowbanned, since shadowbanning is by nature, almost always unconfirmable by the user who is experiencing it. All survey results reported here are significant at the 95% confidence interval.

References

7amleh. (2021). *The Attacks on Palestinian Digital Rights*. The Arab Center for the Advancement of Social Media. https://perma.cc/L6EJ-NMGM

Ahmed, A. (2021, February 11). Facebook Is Widely Rolling Out Its Account Status Tab, Allowing Users To Examine Their Profile, Pages And Groups Restriction History. *Digital Information World*. https://perma.cc/3A6E-39BL

Akpan, P. (2020, November 9). How Social Media Shadowbanning Is Impacting Minority Groups The Most. *Bustle*. https://perma.cc/5P8W-PFXR

Are, C. (2021). The Shadowban Cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 0(0), 1–18. https://perma.cc/7E9A-EATM

Atwood, J. (2011, June 4). *Suspension, Ban or Hellban? Coding Horror*. https://perma.cc/2JQ7-LR2S/

Aziz, F. (2019, November 25). Here is a trick to getting longer lashes! #tiktok #muslim #muslimmemes #islam https://t.co/r0JR0HrXbm [Tweet]. *@ferozaazizz*. https://perma.cc/WDY9-UZCL

Baldauf, J., Ebner, J., & Guhl, J. (2019). *Hate Speech and Radicalisation Online: The OCCI Research Report*. Institute for Strategic Dialogue. https://perma.cc/P9N6-RX3K/

Barrett, P. M., & Sims, J. G. (2021, February). *False Accusation: The Unfounded Claim that Social Media Companies Censor Conservatives*. NYU Stern Center for Business and Human Rights. https://perma.cc/MQ3Y-CZBJ

Bartlett, L. (2021, January 31). How Your Cannabis Company Can Avoid A Shadowban On Social Media. *Forbes*. https://perma.cc/HFF2-69YQ/

Biddle, S., Ribeiro, P. V., & Dias, T. (2020, March 16). Invisible Censorship: TikTok Told Moderators to Suppress Posts by "Ugly" People and the Poor to Attract New Users. *The Intercept*. https://perma.cc/4EVH-JCPN/

Bloch-Wehba, H. (forthcoming). Content Moderation as Surveillance. *Berkeley Technology Law Journal*, 36. https://perma.cc/D3GB-V4AX

Blunt, D., Wolf, A., Coombes, E., & Mullin, S. (2020). *Posting Into the Void: Studying the Impact of Shadowbanning on Sex Workers and Activists*. Hacking//Hustling. https://perma.cc/YP4R-D83X/

Bohn, D. (2017, February 16). One of Twitter's new anti-abuse measures is the oldest trick in the forum moderation book. *The Verge*. https://perma.cc/Y44D-KXYK

Broniatowski, D. A., Jamison, A. M., Johnson, N. F., Velasquez, N., Leahy, R., Restrepo, N. J., Dredze, M., & Quinn, S. C. (2020). Facebook Pages, the "Disneyland" Measles Outbreak, and Promotion of Vaccine Refusal as a Civil Right, 2009–2019. *American Journal of Public Health*, 110(S3), S312–S318. https://perma.cc/PS2M-R58J

Burke, R., Abdollahpouri, H., Mobasher, B., & Gupta, T. (2016). Towards Multi-Stakeholder Utility Evaluation of Recommender Systems. *UMAP*. https://perma.cc/6CZU-EEGR

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512. https://perma.cc/4EBQ-UG7X

Cardi B. (2021, January 4). Twitter release me from twitter shadowban!!! I won't talk about suckin and fuckin nomore .I promise that was 2020 behavior! [Tweet]. *@iamcardib*. https://perma.cc/SP5N-3QXU

Clark-Flory, T. (2019, April 17). *A Troll's Alleged Attempt to Purge Porn Performers from Instagram*. Jezebel. https://perma.cc/XMW4-RK3L

Colombo, C. (2021, August 3). *Kiwi Farms, the forum that has been linked to 3 suicides, was made to troll Chris Chan years before she was arrested on an incest charge*. Insider. https://perma.cc/5986-RVTP

Constine, J. (2019, April 10). Instagram now demotes vaguely 'inappropriate' content. *TechCrunch*. https://perma.cc/JT9L-UV5D/

Cook, J. (2020, February 25). *Instagram's CEO Says Shadow Banning 'Is Not A Thing.' That's Not True*. HuffPost. https://perma.cc/9XH6-CXXZ

Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, 21(4), 895–913. https://perma.cc/Y7TU-7H58

Cotter, K. (2021). "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*. https://perma.cc/74RN-F75K

Coyle, D., & Weller, A. (2020). "Explaining" machine learning reveals policy challenges. *Science*, 368(6498), 1433–1434. https://perma.cc/T6XG-9T98

Davis, J. (2020, October 28). *Instagram changes nudity policy after plus-size model row*. Harper's Bazaar. https://perma.cc/95BQ-WSJC/

Hearing on "Twitter: Transparency and Accountability," United States House of Representatives, 115th Congress (2018) (testimony of Jack Dorsey). https://perma.cc/U9KS-UENJ

Douek, E. (2020, February 11). *The Rise of Content Cartels*. Knight First Amendment Institute. https://perma.cc/3CDJ-BR2R

Duarte, N., Llansó, E., & Loup, A. C. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis*. Center for Democracy & Technology. https://perma.cc/LL8Q-KB55

Duffield, W. (2021, July 22). *Learning about Content Moderation from Ghosts in Virtual Reality*. Cato Institute. https://perma.cc/QYX6-8YN8

Duffy, B. E. (2020). Algorithmic precarity in cultural work. *Communication and the Public*, 5(3–4), 103–107. https://perma.cc/PER2-RTYL

Elliott, V. (2021, January 7). *How vaccine misinformation spreads in Spanish on Facebook*. Rest of World. https://perma.cc/EWT9-5HHH/

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": 33rd Annual CHI Conference on Human Factors in Computing Systems, CHI 2015. *CHI 2015 - Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems*, 153–162. https://perma.cc/74HV-S5X3

Fitzgerald, J., & Sage, J. (2019, June 12). Shadowbans: Secret Policies Depriving Sex Workers of Income and Community. *Tits and Sass*. https://perma.cc/X3UN-9LHK/

Fleming, A. (2021, February 10). The model who made Instagram apologise: Alexandra Cameron's best photograph. *The Guardian*. https://perma.cc/QF9T-T6SY

Flores-Saviaga, C., & Savage, S. (2018, September 12). *Savvy social media strategies boost anti-establishment political wins*. The Conversation. https://perma.cc/H5QU-X9CP

Forsey, C. (2021, July 15). *Instagram Shadowban Is Real: How to Test for & Prevent It*. https://perma.cc/7RQG-P9X3

Gadde, V., & Beykpour, K. (2018, July 26). *Setting the record straight on shadow banning*. Twitter Blog. https://perma.cc/5PUG-GMHN

Garcha, N. (2020, December 19). *Online censorship claims shadow Indian farmer solidarity protests | Globalnews.ca*. Global News. https://perma.cc/23H8-U96X/

Gebel, M. (2020, July 21). Black Creators Say TikTok Still Secretly Hides Their Content. *Digital Trends*. https://perma.cc/ZD9U-8XYW

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12), 4492–4511. https://perma.cc/AG59-ELMW

Global Network Initiative. (2019, January 12). Implementation Guidelines for the Principles on Freedom of Expression and Privacy. *Global Network Initiative*. https://perma.cc/LVL7-LWC8/

Global Network Initiative. (2020). *The GNI Principles At Work: Public Report on the Third Cycle of Independent Assessments of GNI Company Members 2018/2019*. https://perma.cc/P39B-KHFN

Gonzalez, O. (2021, May 26). *Twitch is filling up with hot tub streams: What this trend is all about*. CNET. https://perma.cc/76PA-GLP5/

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1). https://perma.cc/7APU-LCRF

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 466:1-466:35. https://perma.cc/7H2T-AA26

Heldt, A. (n.d.). *Borderline speech: Caught in a free speech limbo?* Internet Policy Review. Retrieved July 27, 2021, from https://perma.cc/4XC4-LAM4

Horten, M. (2021). Algorithms Patrolling Content: Where's the Harm? *SSRN Electronic Journal*. https://perma.cc/7RU4-WT24

Human Rights Watch. (2021, October 8). Israel/Palestine: Facebook Censors Discussion of Rights Issues. *Human Rights Watch*. https://perma.cc/328L-QH9E

Instagram. (n.d.-a). *#boho hashtag on Instagram*. Retrieved February 16, 2022, from https://perma.cc/PJ5C-47F4/

Instagram. (n.d.-b). *#italiano hashtag on Instagram*. Retrieved February 16, 2022, from https://perma.cc/PD99-4CMU/

Instagram. (n.d.-c). *#kansas hashtag on Instagram*. Retrieved February 16, 2022, from https://perma.cc/SMF3-UCVZ/

Instagram. (n.d.-d). *New Features: Instagram Outage Updates And Account Status Inbox | Instagram Blog*. Retrieved November 17, 2021, from https://perma.cc/FQ87-35M2

Instagram. (n.d.-e). *#suicide hashtag on Instagram*. Retrieved February 16, 2022, from https://perma.cc/Q6S5-BQA3/

Instagram. (n.d.-f). *Why aren't Top or Recent posts showing up for a particular Instagram hashtag page?* Instagram Help Center. Retrieved September 10, 2021, from https://perma.cc/GVH4-DMKJ

Instagram. (n.d.-g). *Why can't I search for certain hashtags on Instagram?* Instagram Help Center. Retrieved November 5, 2021, from https://perma.cc/RNZ7-SCX9

Instagram. (2017, February 28). *Instagram for Business*. Facebook. https://perma.cc/6JVQ-AGZJ

Instagram Help Center. (n.d.). *How do I filter out and hide comments I don't want to appear on my posts on Instagram?* Retrieved January 11, 2022, from https://perma.cc/QW8E-Y3A9

Joseph, C. (2019, November 8). *Instagram's murky 'shadow bans' just serve to censor marginalised communities*. The Guardian. https://perma.cc/3YLG-2FFS

Kamara, S., Knodel, M., Llansó, E., Nojeim, G., Qin, L., Thakur, D., & Vogus, C. (2021). *Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems* (p. 38). https://perma.cc/93YH-LXU9/

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction*, 5, 1–44. https://perma.cc/Z2CN-7XAD

Kaushika, P. (2019, February 16). New allegation against Twitter – Modi supporters say PM is 'shadow-banned.' *ThePrint*. https://perma.cc/E49S-S76S/

Kempton, W. (1986). Two Theories of Home Heat Control. *Cognitive Science*, 10(1), 75–90. https://perma.cc/2H28-2KJZ

krispykrackers. (2015, July 28). *On shadowbans*. [Reddit Post]. R/Self. https://perma.cc/YHK8-GHEB

Kumar, S., Cheng, J., Leskovec, J., & Subrahmanian, V. S. (2017). An Army of Me: Sockpuppets in Online Discussion Communities. *Proceedings of the 26th International Conference on World Wide Web*, 857–866. https://perma.cc/3JPT-C8DE

Lakier, G. (2021, July 26). *Informal Government Coercion and The Problem of "Jawboning."* Lawfare. https://perma.cc/Y5LF-SP9C

Le Merrer, E., Morgan, B., & Tredan, G. (2021). Setting the Record Straighter on Shadow Banning. *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 1–10. https://perma.cc/X8Y6-6JCZ

Lee, T. B. (2018, January 13). *Did Twitter engineers just admit to shadow-banning conservatives? Nope*. Ars Technica. https://perma.cc/5VLX-LFEF/

Lorenz, T. (2017, June 7). *Instagram's "shadowban," explained: How to tell if Instagram is secretly blacklisting your posts*. Mic. https://perma.cc/L5BG-Z95Y

Ma, R., & Kou, Y. (2021). "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 429:1-429:25. https://perma.cc/5PFE-CAS4

MarkdownShadowBot. (2018, November 4). *[HOW TO USE] What is this??!* [Reddit Post]. R/CommentRemovalChecker. https://perma.cc/W7H4-RVCW

May, N. (2019, December 5). *Eva Chen on sustainability, shadow banning and the future of Instagram*. https://perma.cc/FAR2-43JQ

McLachlan, S. (2021, August 4). Experiment: I Tried to Get Shadowbanned on Instagram. *Social Media Marketing & Management Dashboard*. https://perma.cc/VZK7-WSUX/

Merlan, A. (2020, August 24). *How Shadowbanning Went from a Conspiracy Theory to a Selling Point*. Vice. https://perma.cc/WB2F-3R7J

Meta Transparency Center. (2022, January 19). *Reducing the distribution of problematic content*. https://perma.cc/3BZY-XU8A/

Midlarsky, M. I., Crenshaw, M., & Yoshida, F. (1980). Why Violence Spreads: The Contagion of International Terrorism. *International Studies Quarterly*, 24(2), 262–298. https://perma.cc/827T-WPBR

Miyaki, K. (2021). *United States Patent: 10994209 - Shadow banning in social VR setting* (Patent No. 10994209). https://perma.cc/RZ5Z-MTTX

Mosseri, A. (2021a, June 8). *Shedding More Light on How Instagram Works*. https://perma.cc/HSY7-6QSW

Mosseri, A. (2021b, August 25). *Breaking Down How Instagram Search Works*. https://perma.cc/X6QR-6GSK

Moynihan, D. P. (1998). *Secrecy: The American experience*. Yale University Press.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11), 4366–4383. https://perma.cc/XZ34-PZZ2

Open States. (n.d.). *shadowbanning—Open States*. Open States. Retrieved February 28, 2022, from https://perma.cc/3A62-HX9E

Oremus, W. (2019a, May 7). How Satan Was Disappeared From Twitter. *OneZero*. https://perma.cc/5XZJ-4JSY

Oremus, W. (2019b, June 3). Twitter Admits It Was Hiding Some People's Tweets by Mistake—Again. *OneZero*. https://perma.cc/PWE3-7PMR

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering Coordinated Networks on Social Media: Methods and Case Studies. *ArXiv:2001.05658 [Physics]*. https://perma.cc/2LB6-XUB3

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

People for the Ethical Treatment of Animals v. Collins, 8:2021cv02413 (United States District Court for the District of Columbia 2021). https://perma.cc/5M7C-5PVX

Pot_Brothers_at_Law. (n.d.). *Pot_Brothers_at_Law (@pot_brothers_at_law) TikTok*. Retrieved January 21, 2022, from https://perma.cc/8DBK-CCJ5?

Pozen, D. (2009). Deep Secrecy. *Stanford Law Review*, Vol. 62, No. 2, p. 257, 2010. https://perma.cc/3E7Q-D3PD

Rao, L. (2013, May 18). The Evolution Of Hacker News. *TechCrunch*. https://perma.cc/Q8EX-C78Q/

Ravi, P. (2021, July 26). *Shadow Bans, Criminal Cases, Takedowns: Inside India's Expanding Digital Crackdown*. Article 14. https://perma.cc/6NNR-QSM2

r/commentremovalchecker. (n.d.). *Reddit API call*. Retrieved February 16, 2022, from https://perma.cc/5RFV-V7FR

Reddit. (n.d.). *R/RedditMartyrs*. Retrieved January 21, 2022, from https://perma.cc/YD38-6MDS/

Reuter, M., & Köver, C. (2019, November 29). Stop complaining about us!: TikTok's Criticism and Competition Guidelines. *Netzpolitik.org*. https://perma.cc/8Y27-M33B/

Richman, J. (2019, October 15). *This is the impact of Instagram's accidental fat-phobic algorithm*. Fast Company. https://perma.cc/5FMQ-XKJN

Robertson, A. (2021, November 12). Meta CTO thinks bad metaverse moderation could pose an 'existential threat.' The Verge. https://perma.cc/U4KJ-MH8G

Ryan, F., Fritz, A., & Impiobato, D. (2020). *TikTok and WeChat: Curating and controlling global information flows*. Australian Strategic Policy Institute. https://perma.cc/24EG-DACU

Sadlier, D. J. (2009). *Latin American Melodrama: Passion, Pathos, and Entertainment.* University of Illinois Press. https://perma.cc/6W8K-X3WF

Salty. (2021, January 23). *Exclusive: An Investigation into Algorithmic Bias in Content Policing on Instagram (PDF download) | Salty*. https://perma.cc/Z58E-WG4Y

Samuels, B. (2018, July 26). *Trump: "We will look into" Twitter for "shadow banning" Republicans* [Text]. TheHill. https://perma.cc/K7HL-TLJB

Serrato, R. (2020, August 31). Detecting Coordination in Disinformation Campaigns. *The Startup.* https://perma.cc/JY7R-ERWN

shadowban.eu. (n.d.). *Twitter Shadowban Test*. Retrieved February 16, 2022, from https://perma.cc/65VC-VBY9/

Shen, H., DeVos, A., Eslami, M., & Holstein, K. (2021). Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 433:1-433:29. https://perma.cc/MT3G-8QFP

Shenkman, C., Thakur, D., & Llansó, E. (2021). *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*. Center for Democracy & Technology. https://perma.cc/P3YF-SB54/

Sheppard, B. (2021). The Reasonableness Machine. *Boston College Law Review*, 62(7), 81. https://perma.cc/KD8E-PF5N

Solon, O. (2021, October 1). *When one pill kills*. https://perma.cc/D423-DFS2/

Soria, L. L. (2020, August 20). *Using A/B testing to measure the efficacy of recommendations generated by Amazon Personalize*. Amazon Web Services. https://perma.cc/U99E-UUUN/

Stack, L. (2018, July 26). What Is a 'Shadow Ban,' and Is Twitter Doing It to Republican Accounts? *The New York Times*. https://perma.cc/LA34-3HW7

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 127:1-127:26. https://perma.cc/WSY6-N3E3

Stepanov, A. (2021, September 23). Sharing Our Content Distribution Guidelines. *About Facebook*. https://perma.cc/W444-LJDV/

Stop Social Media Censorship Act, H 4528, South Carolina Legislature, 124th Session, 2021–2022 (2021). https://perma.cc/7DNY-JGBU

Strauss, E. A., Saif Farooqui, M., SG, Muhammad, R. M., Hwang, M. R., Scheffer, N., & Rhyu, J. (2019). *United States Patent: 10412032 - Techniques for scam detection and prevention* (Patent No. 10412032). https://perma.cc/FFX7-ZVZB

Szakacs, G. (2021, January 18). Hungary mulls sanctions against social media giants. *Reuters*. https://perma.cc/D3Z5-SJW4/

Thompson, A. (2018, July 25). Twitter appears to have fixed "shadow ban" of prominent Republicans like the RNC chair and Trump Jr.'s spokesman. *Vice*. https://perma.cc/6A63-6WRN

TikTok. (2019, August 16). *How TikTok recommends videos #ForYou. Newsroom | TikTok*. https://perma.cc/NX9W-G2EQ

Todd, B. (2013, July 16). *Does Anything Go? The Rise and Fall of a Racist Corner of Reddit*. The Atlantic. https://perma.cc/XR3S-T828/

Triberr. (n.d.). *Triberr Features*. Triberr. Retrieved October 25, 2021, from https://perma.cc/SBQ7-XRA7

Trump, D. (2021, January 6). *"Save America" Rally*. https://perma.cc/ZW5C-46E8

Twitch. (2021, May 21). *Let's Talk About Hot Tub Streams*. Twitch Blog. https://perma.cc/T2KD-WLSQ/

Twitter Help Center. (n.d.). *How to use advanced muting options*. Retrieved January 11, 2022, from https://help.twitter.com/en/using-twitter/advanced-twitter-mute-options

Twitter Transparency Center. (2021). *Removal Requests*. https://perma.cc/6T3C-GYT7

Valens, A. (2020, October 17). *Report says shadowbanning is real—And it's suppressing sex workers*. The Daily Dot. https://perma.cc/NMK7-PS6K/

Van Horne, J. (2020, March 3). Shadowbanning is a Thing—And It's Hurting Trans and Disabled Advocates. *Salty*. https://perma.cc/RZ9K-3AN9/

Vargas, L., Emami, P., & Traynor, P. (2020). On the Detection of Disinformation Campaign Activity with Network Analysis. *ArXiv:2005.13466 [Cs]*. https://perma.cc/Z6LU-NEE2

Vogels, E. a, Perrin, R., & Anderson, M. (2020, August 19). Most Americans Think Social Media Sites Censor Political Viewpoints. *Pew Research Center: Internet, Science & Tech*. https://perma.cc/N32P-XQ5J/

Vogus, C., & Llansó, E. (2021). *Making Transparency Meaningful: A Framework for Policymakers*. Center for Democracy & Technology. https://perma.cc/6ZCN-4Z4Q

Walker, S. (2021, January 14). Poland plans to make censoring of social media accounts illegal. *The Guardian*. https://perma.cc/73XS-83KU

Wisconsin Senate Bill 582, SB 582, Wisconsin Senate, 2021 Regular Session (2021). https://perma.cc/74LD-S5HC

Woolery, L. (2018, March 8). It's All Downsides: Hybrid FOSTA/SESTA Hinders Law Enforcement, Hurts Victims and Speakers. *Center for Democracy and Technology*. https://perma.cc/KW3J-EH3L/

Zhang, M. (2018, March 21). *This New Instagram Shadowban Tester Examines Your Last 10 Posts | PetaPixel*. https://perma.cc/3U7V-PDA4/

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY