Risius, Marten; Blasiak, Kevin Marc

Shadowbanning

Business & Information Systems Engineering

Bus Inf Syst Eng 66(6):817–829 (2024)
https://doi.org/10.1007/s12599-024-00905-3

CATCHWORD

Shadowbanning

An Opaque Form of Content Moderation

Marten Risius · Kevin Marc Blasiak

Received: 21 February 2024 / Accepted: 26 September 2024 / Published online: 28 October 2024

Keywords Shadowbanning · Visibility reduction · Trust and Safety · Content moderation · Algorithmic content moderation · Social media

1 Introduction

Social media platforms face numerous societal, ethical, and political issues. Online extremism (Spiekermann et al. 2022), disinformation campaigns (Starbird et al. 2019), hate speech (Oksanen et al. 2020), and cyberbullying (Chan et al. 2019) are just a few examples of the social media related problems. Social media platforms have responded by implementing various content moderation mechanisms to govern communication on social media platforms (Grimmelmann 2015). Content moderation is a rapidly growing US$ 9.8 Bn market (Bloomberg 2022; Wankhede 2022). Particularly algorithmic content moderation is an increasingly popular approach (Katzenbach 2021). Algorithmic content moderation encompasses platform design decisions that dictate how community members interact with one another and determine who gets to see which content (Duffy and Meisner 2022; Zeng and Kaye 2022). Algorithmic content moderation offers scalable, automated systems that classify user-generated content to inform governance decisions (e.g., removal, geoblocking, account takedown) (Gorwa et al. 2020; Grimmelmann 2015).

The current discussion on content moderation pays little attention to shadowbanning (Gillespie 2022a; Gillespie et al. 2020). Shadowbanning secretly demotes or suppresses visibility of users, content, or groups without alerting the affected entity (Gillespie 2022a). A recent survey of 1,000 U.S. social media users found that about 10% of respondents – typically non-cisgendered, Hispanic, or Republican users report being shadowbanned across all major social media platforms like Facebook, Twitter, Instagram, Reddit, and TikTok (Nicholas 2022). Shadowbanning is the conceptual counterpart to platform's amplification of problematic content for the sake of boosting engagement through recommender algorithms (Gillespie 2022a). Social media platforms generally avoid using the term shadowbanning as part of their content moderation mechanisms. Instead, they refer to it as visibility reduction techniques (Gillespie 2022b). Social media platforms have good reasons to use shadowbanning as it allows them to contain unwanted content without releasing information that would help malicious actors adjust their tactics and avoid detection (e.g., spam bots), to mitigate access to undesirable content (e.g., suicide, pro-eating disorder) (Nicholas 2022), or to avoid polarization and public outcry resulting from certain content moderation decisions (Gillespie 2022a).

Accepted after two revisions by Susanne Strahringer.

M. Risius (✉)
Information Management, University of Applied Sciences Neu-Ulm, Wileystraße 1, 89231 Neu-Ulm, Germany
e-mail: m.risius@business.uq.edu.au

M. Risius
Adjunct Senior Fellow, School of Psychology, University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia

K. M. Blasiak
Center for Technology & Society, TU Wien, Gußhausstraße 27-29, 1040 Vienna, Austria

K. M. Blasiak
Informatics, TU Wien, Favoritenstraße 9-11, 1040 Vienna, Austria

Shadowbanning is also heavily scrutinized, predominantly for its opacity. Shadowbanning prevents users from correcting or disputing content moderation decisions (Nicholas 2022), leaving users left to speculate about whether they have been shadowbanned (Delmonaco et al. 2024) and unclear about the criteria that trigger shadowbanning (Elmimouni et al. 2024). Shadowbanning is accused of systematic bias against minorities (Duffy and Meisner 2022). In a recent content moderation survey that oversampled marginalized identities (i.e., racial and ethnic minorities, LGBTQ + people, trans and/or nonbinary people), 21.78% of respondents reported experiencing shadowbans (Delmonaco et al. 2024). Subjects of shadowbanning report mental and emotional harm (Nicholas 2022) ranging from feelings of frustration, sadness (Delmonaco et al. 2024), marginalization, anxiety, and helplessness (Elmimouni et al. 2024), leading to self-censorship, withdrawal from social media, and financial damages (Delmonaco et al. 2024). The potential ramifications of shadowbanning are expected to extend far beyond the silenced individuals or minority groups directly affected. This mechanism is believed to erode trust and confidence in social media platforms, fostering an environment conducive to conspiracy theories (Chen and Zaman 2024). For instance, shadowbanning fuels beliefs that platforms hold biased agendas, such as aligning with specific governments (e.g., "platforms align with the state of Israel"). Similarly, shadowbanning can exacerbate societal polarization by filtering certain opinions or individuals from public discourse. This can bias the process of forming public opinion, for example, when restricting pro-Palestinian voices on Facebook during the Israel-Hamas war (Elmimouni et al. 2024) or TikTok's suppression of #BlackLivesMatter and LGBTQ + content (Delmonaco et al. 2024). Finally, shadowbanning can be weaponized by malicious actors to silence dissenting voices (Nicholas 2022), undermining open dialogue and empowering those who seek to manipulate online discourse.

A balanced and informed discussion of shadowbanning is urgently needed. Related research is still nascent and – to the best of our knowledge – absent in the field of Information Systems (IS). The objective of this article is to introduce IS practitioners and researchers to shadowbanning. By introducing shadowbanning, we aim to make three key contributions. First, we contribute to the emergent literature that raises awareness for this opaque content moderation mechanism (Gillespie 2022a). Shadowbanning complements existing IS research on content moderation beyond the more commonly discussed forms of annotating (He et al. 2024; Kim and Dennis 2019; Kim et al. 2019), banning (Russo et al. 2023), blocking (McDonald 2022), and deplatforming (Keller 2019). But also other related IS research on platform governance (Halckenhaeusser et al. 2020), algorithmic audiencing (Riemer and Peter 2021), and algorithmic control (Benlian et al. 2022) ought to consider the role of algorithms in secretly demoting content – in addition to their augmenting, amplifying, and serendipitous effects (Milli et al. 2023). Second, we offer conceptual clarity into what constitutes shadowbanning. Building a common understanding of shadowbanning ought to help bridge the gap between social media platforms who avoid the term shadowbanning (Gillespie 2022b), users who speculate whether they have been subjected to this practice (Elmimouni et al. 2024), researchers who try investigate this phenomenon (Jaidka et al. 2023), and lawmakers who need to understand this mechanism to devise meaningful regulation (Nicholas 2023). Third, we outline ways in which information systems research with its focus on sociotechnical systems can help address and inform the related conversation around content moderation, censorship, freedom of expression to contribute to society and make the online environment safer for everyone (Sarker et al. 2019; Spiekermann et al. 2022).

2 Background

2.1 Definition and Origins of Shadowbanning

The term shadowbanning first appeared in 2001, where it referred to the mechanism of removing posts for everyone else except for the poster in an online forum (Savolainen 2022). It reached broader public awareness in 2018 when US conservatives began accusing Twitter (now X) of shadow-banning 'prominent Republicans' by not suggesting them on the platform's autofill drop-down search bar (Savolainen 2022; Stack 2018). Shadowbanning trades under many different names such as stealth banning, ghost banning, hell banning, comment ghosting, visibility moderation, (visibility) reduction, suppression, borderline content policies, and uninformed or undisclosed content moderation (Gillespie 2022a, 2022b; Jaidka et al. 2021; Nicholas 2023). Shadowbanning suppresses the visibility or reach of content, users, or groups without alerting the affected party (Gillespie 2022a). Platforms limit the conditions under which the content circulates, for example, whether it appears as a recommendation, a search result, in a news feed and users' queries, or in a stream of comments (Gillespie 2022b). It is an opaque mechanism where content remains theoretically accessible and visible to (some) users (e.g., the poster themselves) (Savolainen 2022). It is important to emphasize that today's understanding of shadowbanning by demoting people, groups, or content with algorithmic support is much more nuanced than the original 2001 mechanism of simply withholding content from anyone (Gillespie 2022b).

Table 1 Summative overview of different types of shadowbans

Shadowban type	Description	Outcome	Examples from literature
Ghost bans	Content is hidden from everyone except the account holder	Complete invisibility to others	Jaidka et al. (2023)
Search bans	Content is removed from the platform's internal search index	Reduced discoverability, contained visibility	Goldman (2021)
Search suggestion bans	Content is removed from the platform's search "auto-suggest" feature	Decreased visibility, potential loss of organic reach	Goldman (2021)
Downtiering	Content's visibility is reduced or limited	Decreased visibility, reduced engagement	Jaidka et al. (2021); Jaidka et al. (2023); Nicholas (2022); Ryan et al. (2020)

Four main types of shadowbans (see Table 1) can be identified (Jaidka et al. 2023): (1) *Ghost bans* are the most restrictive type of shadowbans that obfuscate content by allowing only the accountholders themselves to see their own content (Jaidka et al. 2023). (2) *Search bans* prohibit the intended encounter with users or content by removing them from the platforms' internal search index (e.g., quarantined subreddits from Reddit's internal search, false posts from Instagram's hashtag search) (Goldman 2021). (3) *Search suggestion bans* remove content from the platform's internal search engine's "auto-suggest" feature (Goldman 2021) so that the account does not appear when others look it up in the search interface (Jaidka et al. 2023). (4) *Downtiering* deprioritizes content to limit the unintentional exposure and to make it unlikely for other users to find it by hiding replies under an interstitial and only loading when prompted (Jaidka et al. 2021, 2023; Nicholas 2022; Ryan et al. 2020), downgrading the internal search visibility, reducing internal promotion (e.g., recommendations), and reducing or removing navigation links (e.g., explore pages) (Goldman 2021).

Shadowbanning differs from the standard content curation mechanisms employed by social media platforms through recommender algorithms. Content recommendation follows a different set of concerns and priorities (Gillespie 2022a). Recommender systems and newsfeeds *select for* what is deemed most appealing to optimize for engagement measured by the time spent on the platform, the types of actions taken, and satisfaction proxies (Bucher 2018; McKelvey and Hunt 2019). They collect and analyze user data together with the corpus of all available content to personalize user feeds and maximize user engagement. Content that is more positive – e.g., in terms of recency, serendipity, close tie popularity – has greater likelihood to be selected by recommender algorithms. Shadowbans are not simply an outcome or byproduct of recommender algorithms because shadowbanned content is actively suppressed or obscured, even in search results (Jaidka et al. 2023). Shadowbanning aims to *select out* - what is deemed least appealing based on negative signals that indicate an item ought not to be recommended (Gillespie 2022a). The gold standard for Trust & Safety performance is 'prevalence' and shadowbanning aims to minimize the function of how often users view certain content (Fishman and Harris 2023).

We broadly understand shadowbanning as social media platforms' use of algorithms to align user behavior, accounts, and content with organizational objectives. This conceptualization closely resembles the prevalent definition of algorithmic control (Wiener et al. 2023), being embedded into a broader organizational context (Alizadeh et al. 2023) with human controllers (Cram and Wiener 2020) and organizational intentions (Kellogg et al. 2020; Sullivan et al. 2024), fulfilling sanctioning functions – among others – (Hirsch et al. 2023), complementing and relieving but not entirely replacing human actors (Wiener et al. 2023). Current research on algorithmic control functions (Alizadeh et al. 2023; Hirsch et al. 2023; Kellogg et al. 2020; Sullivan et al. 2024) is yet to consider visibility reduction techniques from platforms like Uber (Uber 2023). Shadowbanning's implicit form of algorithmic control delivery by avoiding to alert users also appears to be an extreme form of the previously investigated algorithmic opacity and limited transparency of algorithms (Möhlmann et al. 2023). Given that algorithmic management researchers recognize the relation of algorithmic control to broader organizational contexts beyond work settings (Cameron et al. 2023), we draw on the related algorithmic management and control literature (Benlian et al. 2022) to guide our conceptualization of shadowbanning in the following (Fig. 1).

In the context of shadowbanning, the human controllers are Trust & Safety teams who define and communicate organizational intentions in the form of platform policies or policy violations, help implement policies into content moderation algorithms, manage or conduct content

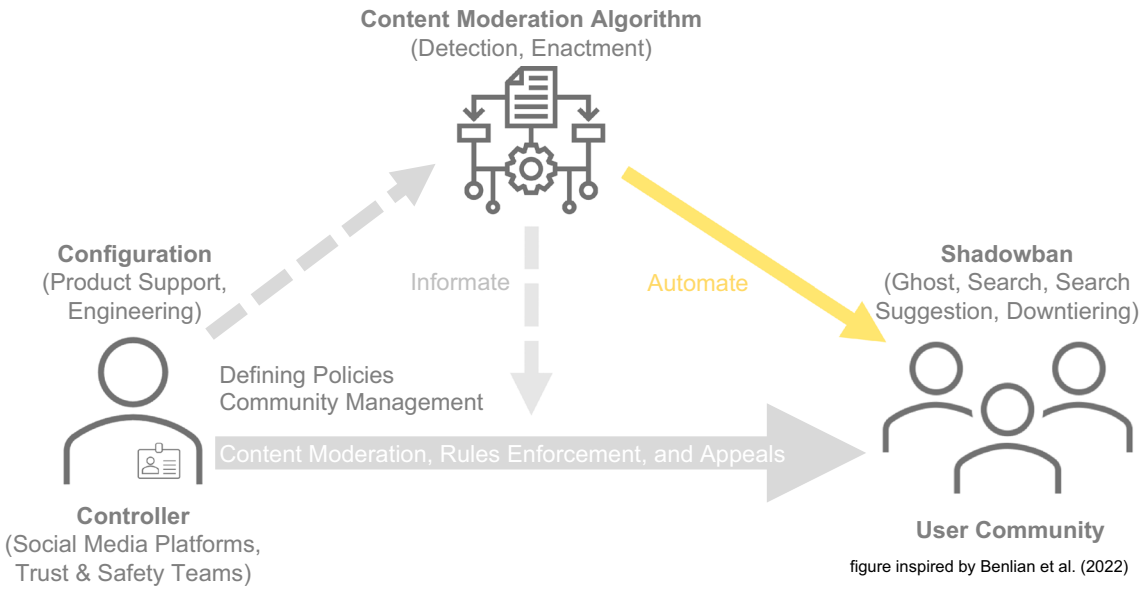

Fig. 1 Conceptualizing shadowbanning as a form of algorithmic control

moderation themselves, and deal with content moderation appeals (Fishman 2023). Shadowbans are typically automated (e.g., to reduce visibility for spam bot posts) but still allow human moderators to enforce shadowbans directly (e.g., to occult certain types of users) (Biddle et al. 2020; Duffy and Meisner 2022). Algorithmic content moderation automates the human Trust & Safety teams moderation tasks (He et al. 2024) in order to gain scalability and consistency in content moderation (Jiang et al. 2023). For shadowbanning this involves the algorithmic detection of unwanted content or users based on Trust & Safety policies and enactment of respective ghost-, search-, search suggestion ban, or downtiering. Shadowbanning solutions can be developed by platforms themselves or sourced from external entities (e.g., Reddit's AutoModerator) (Wright 2022).

2.2 The Role of Shadowbanning as a Form of Content Moderation

Shadowbanning is part of social media platforms broader content moderation efforts to contain the prevalent online issues which means '*governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse*' (Grimmelmann 2015, p. 47). While content moderation mechanisms are hypothetically unlimited from a technological standpoint, research has identified a range of common mechanisms (Goldman 2021) (Table 2). These content moderation mechanisms vary (1) in their degree of severity and (2) regarding their underlying philosophy.

The degree of severity is commonly broken up into either "hard" or "soft" types of content moderation (Zannettou 2021). Hard moderation means to suspend, block or remove content or entities from social media platforms (Gorwa et al. 2020). Soft forms of content moderation warn about content or contain its impact without suspension or take-downs (Jaidka et al. 2023). Content moderation mechanisms' underlying philosophies differ in the degree to which they rely on nurturing or punishing to create a positive online environment (Jiang et al. 2023). Punishing is a reactive approach that primarily focuses on applying consequences for rule-violating behaviors. Nurturing predominantly intends to educate, improve or reform online (mis)behavior. In the following, we apply these dimensions to integrate the different forms of shadowbans into the broader context of other content moderation mechanisms as identified and described by Goldman (2021) (Table 2).

From a philosophical standpoint, shadowbanning is particularly opaque compared to nurturing mechanisms. Users are usually not informed that (or for how long) they are shadowbanned or need to retrieve the information themselves when possible (Silva 2022). Indeed, a major reason for using shadowbans is to disallow malicious actors (e.g., spam bots) to learn from and adjust to content moderation algorithms (Biddle et al. 2020). This prevents users from learning from the shadowbanning decisions and is thus a more punitive measure rather than one that nurtures user behavior in line with community standards.

The different forms of shadowbans vary in their degree of severity (Jaidka et al. 2023). Shadowbans can include both hard and soft types of content moderation depending

Table 2 Conceptualization of shadowbanning among other content moderation mechanisms

		Philosophy	
		Nurturing	Punishing
Degree of Severity	Hard	Assign strikes/warnings	Fine author/impose liquidated damages Outing/unmasking
		Community service	Put user/content on industry-wide blocklist
		Educate users	Remove content
		Redirect method††	Report to law enforcement
		Restorative justice/apology	Suspend account
			Suspend content
			Terminate account
			Ghost ban†
			Remove from internal search index*
	Soft	Age-gate	Disable comments
		Counterspeech	Edit/redact content
		Display content only to logged-in readers	Forfeit accrued earnings
		Interstitial warning	Nofollow authors' links
		Shaming	Outing/unmasking
		Suspend future earnings	Reduce service levels
		Terminate future earnings	Reduced virality
		Warning legend	Relocate content
			Remove credibility badges
			Remove from external search index
			Suspend posting rights
			Downgrade internal search visibility***
			No auto-suggest**
			No/reduced internal promotion***
			No/reduced navigation links***

Types of shadowbans: *search ban, **search suggestion ban, ***downtiering; adaptations from Goldman (2021): †originally referred to as "shadowban", ††adopted from Scrivens and Gaudette (2024); in non-remedial contexts (e.g., remove from external search index, nofollow authors' links) (Goldman 2021) the classification of the punishing philosophy occurs based on the mechanisms' reactive nature

on the degree to which they enforce visibility restrictions (i.e., do not recommend at all/ as much/ to some) (Gillespie 2022a). A ghost ban matches an account suspension without notifying the user (Goldman 2021). Similarly, shadowbans that severely restrict visibility in situations with a short "lifespan" of engagement equate to a suspension (Jaidka et al. 2023) (e.g., removal from the internal search index). Instead of making content inaccessible, shadowbans can also greatly conceal items to make it less likely that users may find them (Jaidka et al. 2021; Nicholas 2022; Ryan et al. 2020) (e.g., downgrade internal search visibility, no auto-suggest in the search function, no or reduced internal promotion, no or reduced exposure in navigation links such as "most popular" or "newly available"). Overall, we consider shadowbanning to be a predominantly punitive mechanism with its specific forms differing in their degree of severity between hard and soft forms of content moderation.

2.3 Unpacking the Controversy Surrounding Shadowbanning

The debate around the prevalence and impact of shadowbanning is highly politicized. On one side, social media platforms are reluctant to admit to shadowbanning (Gillespie 2022a), often framing it as a specific kind of behavior (e.g., "*it doesn't hide people's content for posting too many hashtags*") (Cotter 2021, p. 1234) or referencing its aforementioned original meaning in internet forums (Gillespie 2022b; Savolainen 2022). The absence of a unified, industry-accepted definition for shadowbanning (Gillespie 2022a) complicates discussions and fuels misunderstandings (Elmimouni et al. 2024). Platforms are understandably wary of being scrutinized for their policies (either for being interventionist and biased, or opaque and unaccountable) and aim to avoid the politicization of their content moderation practices (Gillespie 2022a).

From the platforms' perspective, shadowbanning serves as an effective tool to manage problematic content. First, it helps mitigate "lawful but awful" content. This content that trades under different names across platforms (e.g., borderline, sensitive, harmful, undesirable, or objectionable content) (Cotter 2021) is considered detrimental to the user experience, threatens the health of the community, is misleading or salacious. It fails to violate platform policies (Gillespie 2022b) and therefore evades other forms of content removal. Examples include pro-eating disorder), sexually suggestive, firearm, and implicitly drug-related content (Gillespie 2022a) or misinformation (Gillespie 2022a). Second, shadowbanning allows platforms to hide moderation tactics from malicious actors, such as spammers (e.g., clickbait, links to malicious or deceptive sites) (Cotter 2021), bots (e.g., astroturfing, sockpuppeting), or orchestrated disinformation campaigns, preventing them from adjusting their strategies (Llewellyn et al. 2019; Nicholas 2022). Third, it provides a means to avoid public outcries over "censorship" while maintaining flexible moderation, particularly in response to evolving threats like increasing radicalization (Gillespie 2022a).

On the other side, the lack of transparency around shadowbanning has fueled folk theories and speculation among users, policymakers, and watchdogs (Jaidka et al. 2023) (e.g., Elon Musk uses shadowbanning on X to suppress Tesla's employee union account (Masnick 2023)). Reports, especially from marginalized communities, suggest shadowbanning is used to suppress certain voices (Elmimouni et al. 2024). User surveys and interviews on shadowbanning are often discarded through "black box gaslighting" by asserting that users do not have sufficient understanding of content curation algorithms to determine whether they were shadowbanned (Cotter 2021). Platforms explain media reports on shadowbanning as technical glitches, the users' failure to create engaging content, or as a matter of chance through the platform's black-box algorithms (Cotter 2021). This back-and-forth erodes public trust in social media platforms and contributes to societal polarization (Chen and Zaman 2024; Jaidka et al. 2023).

The scrutiny surrounding shadowbanning also stems from its potential for misuse. Cases have emerged where shadowbanning was allegedly used to meet government demands (e.g., silence dissent in China, contain Covid-19 messages in the USA) (AP 2024; Hern 2019), marginalize minority voices (e.g., handicapped, black, or LGBTQ + users) (Delmonaco et al. 2024; Duffy and Meisner 2022), or manipulate public opinion (e.g., mute pro-Palestinian posts) (Chen and Zaman 2024; Elmimouni et al. 2024; Luu 2023). This form of exclusion can result in emotional distress (Lutz and Schneider 2021), financial harm for content creators (Duffy and Meisner 2022), and increased vulnerability to online attacks (for an overview, see Delmonaco et al. 2024; Nicholas 2022). For marginalized groups, shadowbanning can lead to their exclusion from public discourse or bans by association (Delmonaco et al. 2024; Elmimouni et al. 2024). On a broader societal level, the secrecy and lack of recourse foster distrust and conspiracy theories, while entrenching societal divides (Chen and Zaman 2024; Nicholas 2022).

2.4 Establishing the Prevalence of Shadowbanning

The opaque nature of shadowbanning has inspired considerable work on establishing the prevalence of shadowbanning. A major source of evidence are social media platforms themselves when reporting content moderation mechanisms that impose a form of visibility reduction while avoiding the term "shadowbanning" (Merrer et al. 2021). Reddit is the only platform that openly confirms the traditional form of shadowbans (Nicholas 2022). Elon Musk promised to provide more transparency on shadowbanning after taking over Twitter (now X), which is yet to be delivered (Perez 2023) and instead faces scrutiny himself over shadowbans on posts with links to other social media platforms (Newton 2023). Meta's Facebook and Instagram were the first major social media platforms that declared pursuing ways to algorithmically reduce user engagement with borderline content in May 2018 (Gillespie 2022a; Zuckerberg 2021). YouTube, X, LinkedIn, and TikTok have since disclosed applying similar strategies to dealing with sensitive content (Duffy and Meisner 2022) and Uber reports visibility reductions for restaurants in response to unfavorable customer reviews (Uber 2023) (for more details on platform statements on shadowbanning, see Delmonaco et al. 2024; Gillespie 2022a).

Beyond official social media platform statements concerning visibility reduction mechanisms, additional evidence suggests the existence and prevalence of shadowbanning. Anecdotal evidence comes from trace ethnography, surveys or interviews (Nicholas 2022; Wright 2022), analyses of platforms' content moderation patents (Nicholas 2023), and users who closely monitor their engagement statistics (e.g., content creators) (Duffy and Meisner 2022; Zeng and Kaye 2022). Other evidence comes from sources that are difficult to verify such as internal whistle-blowers (Chen 2019), information leaks (Gillespie 2022a), or investigative journalism (Colve 2018; Merlan 2020). To determine the extent of shadowbanning, a recent survey of 1,006 social media users found that 9.2% report having been shadowbanned. Of these 8.1% were on Facebook, 4.1% on Twitter (now X), 3.8% on Instagram, 3.2% on TikTok, 1.3% on Discord, 1% on Tumblr, and less than 1% on YouTube, Twitch, Reddit, NextDoor, Pinterest, Snapchat and LinkedIn (Nicholas 2022). A content

moderation survey that oversampled marginalized identities (i.e., racial and ethnic minorities, LGBTQ + people, trans and/or nonbinary people) even found that 21.78% of respondents reported shadowbanning (Delmonaco et al. 2024).

Recently, researchers have begun collecting statistical evidence for the presence and extent of different forms of shadowbanning. Analyses of 41 k to over 2.5 million Twitter (now X) accounts found that between 3–6.2% of accounts had been shadowbanned at least once (Jaidka et al. 2021; Merrer et al. 2021). These studies identify characteristics that increase the likelihood of shadowbans. Some of these predictors are particularly new accounts (less than two weeks old) with low follower numbers (below 200), using incivility (negative or offensive terms) or posting pictures without text messages, and displaying bot-like behavior (high botometer score) (Jaidka et al. 2023; Merrer et al. 2021). A verified account (e.g., the blue checkmark on X) helps to drastically reduce the chance of a shadowban (Jaidka et al. 2023; Jorgenson 2022). Based on these findings, computer scientists were then able to develop tools or workarounds that identify whether accounts are shadowbanned for different platforms like X (hisubway,[1] yuzurisa[2]) or Reddit (r/CommentRemovalChecker, r/ShadowBan) (Nicholas 2022).

3 Challenges and Opportunities for IS Research

Research on shadowbanning mechanisms and their implications remains in its early stages. While legal scholars, communication researchers, and computer scientists have contributed to this nascent field, existing studies primarily rely on qualitative methods. These methods often involve interviews with affected content creators or utilize anonymous sources within social media companies mechanism (Cotter 2021; Savolainen 2022; Zeng and Kaye 2022) or the analysis of content moderation patents (Nicholas 2023). Information systems scholars can address shadowbanning to help resolve these apparent issues and bridge the conversation between social media platforms, users, and regulators. Applying a sociotechnical perspective, information systems research can help mitigate the human and societal implications of this form of algorithmic content moderation and facilitate productive discourse on this challenge between the stakeholders (i.e., businesses, governments, NGO's) (Gorwa 2022). Building upon the conceptualization of shadowbanning as an algorithmic control process involving various entities (user community, social media platforms' Trust & Safety Teams, content moderation algorithms; Fig. 1), the following sections will explore a non-exhaustive list of illustrative research questions (Table 3).

3.1 User Community

While there is currently little anyone can do if their account or content is shadowbanned, researchers identified common strategies that creators use to prevent and cope with shadowbans: Suppression, experimentation, circumvention, and resignation (Duffy and Meisner 2022). Similarly, creators' experiences helped to develop techniques for individual users to detect whether they were shadowbanned (Columbres 2023). However, user reports of shadowbanning frequently encounter "black box gaslighting," a phenomenon where platforms dismiss these concerns by claiming users misunderstand the complexities of content moderation algorithms (Cotter 2021). IS research can explore methods for users to safely communicate potential shadowbanning experiences (1.1)? Similarly, shadowbanning has significant implications for the targets (Myers West 2018; Nicholas 2022). Core Self-Evaluation theory (Bono and Judge 2003; Judge et al. 2003), for example, can help us understand (and mitigate) the shadowbanning effects on users' self-esteem, self-efficacy, locus of control, and emotional stability (1.2). Lastly, shadowbanning has considerable monetization ramifications for creators (Duffy and Meisner 2022; Myers West 2018). Self-determination theory (Ryan and Deci 2000) allows to examine the complexities of the algorithmically driven workplace at the individual level (Benlian et al. 2022). Here it can help us understand how shadowbanning affects content creators or regular users' (e.g., job candidate) perceived competence, autonomy, and relatedness after being shadowbanned (1.3). Shadowbanning also poses a major challenge for marginalized groups that appear to be disproportionally targeted (e.g., "ugly, poor, or disabled" users, "Black Lives Matter" filter, LGBTQ + keywords) (Duffy and Meisner 2022; Ryan et al. 2020; Walsh 2022). The DIME model (Louis et al. 2020) allows IS scholars to explore the responses by these movements to shadowbanning (1.4).

3.2 Social Media Platforms (Trust&Safety Teams)

Shadowbanning helps platforms fulfil important societal tasks by moderating various forms of harmful content. While its secrecy is strongly contested (e.g., promote conspiracies, reduce trust) (Nicholas 2022), shadowbanning's opaque nature helps to avoid politicizing content moderation (Gillespie 2022a). The broader implications of content moderation mechanisms are controversially discussed among law and policy scholars (e.g., Douek (2022); Goldman (2021); Gorwa et al. (2020)), in communications

[1] https://hisubway.online/shadowban/.

[2] https://shadowban.yuzurisa.com.

Table 3 Suggested areas of research to advance knowledge on shadowbanning

Topic		Proposed research questions
User community	1.1	How can people safely communicate that they were shadowbanned?
	1.2	How does shadowbanning affect users' self-esteem, self-efficacy, locus of control and emotional stability?
	1.3	What effects has shadowbanning on users' perceived competence, autonomy, and relatedness?
	1.4	How do marginalized groups respond to being shadowbanned?
	1.5	How does shadowbanning influence users' online behavior, such as self-censorship or avoidance of certain platforms?
Social media platforms (Trust & Safety Teams)	2.1	How do shadowbanning's trade-offs compare to other forms of content moderation?
	2.2	How effective is shadowbanning in reducing polarization and protecting free speech compared to other forms of content moderation?
	2.3	What factors drive platform's exercise of shadowbanning?
	2.4	What are the brand safety implications of shadowbanning for advertisers?
	2.5	What characterizes ethically acceptable shadowbanning cases?
	2.6	How to design shadowbanning reporting structures that offer reasonable amounts of transparency?
	2.7	What other types of platforms use shadowbanning to moderate content?
Content moderation algorithm	3.1	How can shadowbanning be detected across platforms?
	3.2	How can advancements in artificial intelligence and machine learning improve the accuracy and fairness of shadowbanning algorithms?
	3.3	How can we implement algorithmic sensemaking into shadowbanning decisions to reduce algorithm aversion?
	3.4	What are the implications of shadowbanning's reductionist effects for free speech and algorithmic audiencing?
	3.5	What situations prompt manual vs automated shadowbanning?
	3.6	What are the trade-offs regarding efficacy and ramifications between different types of shadowbans?

research (e.g., Suzor et al. (2019)), and computer scientists [e.g., Seering (2020), Jhaver et al. (2019a), Jhaver et al. (2019b)]. IS research should join the debate and explore shadowbanning's ability to mitigate societal or group polarization (Sunstein 2018) or protect free speech (Riemer and Peter 2021) compared to other forms of content moderation (2.1, 2.2). Alternatively, research demonstrates that a verified account and blue checkmarks are protective factors against shadowbanning (Jaidka et al. 2023; Jorgenson 2022), which are now partly offered for sale (Klar 2023). IS researchers could address this issue, for example, by applying the critical theory of power and ethics by Foucault (2007) to explore whether shadowbanning is an ethical, political, or primarily a technical problem (Siapera and Viejo-Otero 2021). Researchers could also aim to delineate the intended versus unintended consequences of systematic biases within shadowbanning (2.3).

Moderating problematic social media content is a major concern for advertisers that want to avoid having their ads displayed next to hateful content (Cooban 2023). While shadowbanning confines the visibility of problematic content, it remains potentially accessible online. IS scholars can apply, for example, brand safety (Bishop 2021) or situational crisis communication theory concepts (Coombs 2007) to assess advertisers' concerns regarding platforms shadowbanning (2.4). We highlighted the heated politicized debate around the term shadowbanning (Cotter 2021; Gillespie 2022a). The good reasons for performing shadowbans are often met with harsh criticism of censorship and marginalization. Researchers could follow other examples of IS research on contested online behaviors (e.g., doxing) (Franz and Thatcher 2023) to determine boundary conditions for ethically acceptable shadowbanning (2.5) and whether some platforms' efforts that allow users to trace their account status means meaningful improvements (2.6) (Silva 2022; Zakharchenko 2024). In this regard, while we have summarized the evidence on the prevalence of shadowbanning on social media platforms, other platforms (e.g., Gig-economy) also report forms of visibility reduction (Uber 2023). Shadowbanning offers research the opportunity to expand their conceptualization of algorithmic control functions (Alizadeh et al. 2023; Hirsch et al. 2023; Kellogg et al. 2020; Sullivan et al. 2024), refine algorithmic management to benefit society at large (Möhlmann et al. 2023), and explore the prevalence

of shadowbanning on other (non-social media) platforms (2.7).

3.3 Content Moderation Algorithm

Shadowbanning is primarily implemented through content moderation algorithms. These algorithms automate the large-scale enforcement of human-designed Trust & Safety policies (He et al. 2024; Jiang et al. 2023). Reliably detecting shadowbanning remains a key challenge due to its opaque nature and the platform's unwillingness to share deeper insights into their algorithms. It requires identifying normal or expectable levels of engagement given the characteristics of the user (e.g., number of followers) and their content (e.g., topic, hashtags, links) (Gillespie 2022b; Nicholas 2022). Some detection tools and methods exist (e.g., hisubway, yuzurisa, r/CommentRemovalChecker) (Jaidka et al. 2021; Merrer et al. 2021), however, they frequently defunct (e.g., formerly Treiberr, shadowban, whosban) (Nicholas 2022) and are limited to select platforms. IS scholars could leverage, for example, new APIs for Meta (Ryan-Mosley 2023) or TikTok (TikTok 2024) and publicly available datasets (SOMA 2024) to develop new shadowbanning detection tools (3.1). Given the prevalence of lawful but awful content across internet platforms such as Spotify (Lima and Schaffer 2022) and Amazon (Bogle 2022; Dreisbach 2021), IS researchers could thereby contribute valuable insights into across platforms shadowbanning differences (e.g., Uber, Deliveroo, Craigslist). In a similar vein, the pervasive issues of algorithmic biases against vulnerable user groups (Feuerriegel et al. 2020; Spiekermann et al. 2022) and lack of transparent, explainable, or interpretable algorithms (Kim and Routledge 2018) also affects algorithmically enforced shadowbanning decisions (Cotter 2021). Prior research has established, for example, that verified accounts and higher engagement numbers are a protective factor against shadowbanning (Jaidka et al. 2023). Biases that reward celebrity status, audience size, and the ability to pay for account verifications challenge the notion of social media platforms as means for crowd empowerment (Leong et al. 2019). IS scholar could use the aforementioned resources to refine the detection algorithms and help debias and elucidate shadowbanning decisions or follow the existing precedent interview and case-studies (Delmonaco et al. 2024; Elmimouni et al. 2024) to better understand algorithmic bias in the context of shadowbanning (3.2).

The algorithmic management and control literature (Benlian et al. 2022; Cameron et al. 2023) proposes users' algorithm aversion against automatically enforced decisions (Berger et al. 2021; Spiekermann et al. 2022). Targets of shadowbanning loose trust into platforms and regulators giving rise to conspiracy theories (Nicholas 2022). Targets of shadowbans report considering leaving the platforms or withdrawing from social media altogether (Delmonaco et al. 2024; Elmimouni et al. 2024). IS scholars could build on the respective algorithm control and sensemaking literature (Möhlmann 2021; Möhlmann et al. 2021; Möhlmannn et al. 2023), for example, to design shadowbanning moderation mechanisms that are perceived as less threatening. This would also contribute insights into the general role of algorithms on trust building and the individual's attitudes towards algorithmic control (Kizilcec 2016; Lee 2018) and designing control algorithms in a way that benefits platforms and society at large (Cameron et al. 2023) (3.3). Similarly, the emergent literature on algorithmic audiencing recognizes the role of algorithmic content moderation and distribution for free speech (Riemer and Peter 2021). This new, extended understanding of free speech in social media could be complemented by a greater consideration of the reductionist effects of shadowbanning through recommender systems – as opposed to the predominant focus on content virality and amplification (3.4) (Milli et al. 2023). Lastly, while shadowbanning is predominantly algorithmically-enforced, it has been found to be enacted by humans. IS researchers could engage with Trust & Safety teams to learn more about the differences between manual and automated shadowbanning processes and provide much needed insights into the process of content moderation more broadly (3.5) (Gorwa et al. 2020; Grimmelmann 2015). Such research could lead to more dedicated research on the different forms of shadowbanning (i.e., ghost, search, search suggestion, downtiering) (Jaidka et al. 2023) in terms of their efficacy, ramifications, and fields of application. We believe a more nuanced understanding of the different forms of shadowbans is essential to inform public discourse (3.6).

4 Outlook

Social media platforms use algorithms to control user attention, a key resource in our increasingly digital world (Zeng and Kaye 2022). A lot has been written on how these algorithms are designed to maximize user engagement, promoting controversial or provocative content on the fringes of mainstream discourse (Zuckerberg 2021) and the question of whether these algorithms induce societal polarization (Bakshy et al. 2015; Guess et al. 2023; Robertson et al. 2023), promote online extremism (Risius et al. 2024), or form filter bubbles and echo chambers (Bruns 2021). Meanwhile, the opposite use of algorithms to demote, hide, and reduce the visibility of content is mostly disregarded (Gillespie 2022a). Shadowbanning reduces or suppresses the visibility and reach of content, users, or groups without notifying the affected party. These

algorithms enable platforms to obscurely organize content by demoting and hiding content or users instead of advertently blocking or deleting them (Merrer et al. 2021). Shadowbanning has been found to disadvantage marginalized groups and has severe ramifications for individuals, communities, and society (Nicholas 2022). This article aims to change the outlook for shadowbanning in three regards.

First, given its opaque character, we aim to raise awareness for the issue of shadowbanning among the public, researchers, and regulators. We argue that shadowbanning should be part of the current conversations around (algorithmic) content moderation (Gorwa et al. 2020; Grimmelmann 2015), algorithmic audiencing and free speech (Riemer and Peter 2021), algorithmic biases (Spiekermann et al. 2022), and algorithmic control (Benlian et al. 2022). *Second*, users who report shadowbanning are often met with "black box gaslighting" (Cotter 2021). This article compiles various forms of evidence for the prevalence of shadowbanning. We therefore join calls to move past the red herring question whether or not shadowbanning exists (Gillespie 2022a, 2023; Nicholas 2023). While we recognize the importance of developing ways to detect shadowbanning, we need to expand the focus on its societal implications, ethical considerations of (non)acceptable shadowbanning, and its inherent trade-offs (e.g., between transparency vs. opacity, level of user activity vs. quality of content, or nurturing vs. punishing content moderation) (Jiang et al. 2023). *Third*, shadowbanning lacks conceptual clarity which facilitates the politicization and conspiratorial theorizing (Gillespie 2022b; Nicholas 2023). This allows platforms to misconstrue and then deny shadowbanning (Cotter 2021). It also allows lawmakers to use shadowbanning for weaponizing free speech regulation (e.g., current supreme court case *Moody v. NetChoice, LLC)* (Nicholas 2023). We have witnessed the politicization and weaponization of other insufficiently defined issues such as fake news before (Kaye 2019). Acc ordingly, some experts argue to abandon the term shadowbanning altogether and move to a less contended term (e.g., visibility reduction or undisclosed content moderation) (Gillespie 2022b; Nicholas 2023). However, given the great public awareness for the issue, we are of the opinion that scientists ought to remain part of the conversation and offer scientific insights. Hence, we aim to offer conceptual clarity on what constitutes shadowbanning and hope to inspire more dedicated research to inform public debate.

Acknowledgements Marten Risius receives generous funding support from Bavarian State Ministry of Science and the Arts through the Distinguished Professorship Program as part of the Bavarian High-Tech Agenda. He is the recipient of an Australian Research Council Australian Discovery Early Career Award (project number DE220101597) funded by the Australian Government. We greatly appreciate Ka Yan Sabrina Ng for inspiring us to investigate shadowbanning.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

AP (2024) Zuckerberg says the White House pressured Facebook to 'censor' some Covid-19 content during the pandemic. https://www.pbs.org/newshour/politics/zuckerberg-says-the-white-house-pressured-facebook-to-censor-some-covid-19-content-during-the-pandemic. Accessed 26 Sep 2024

Alizadeh A, Hirsch F, Jiang J, Wiener M, Benlian A (2023) A taxonomy of algorithmic control systems. In: 44th International Conference on Information Systems, Hyderabad, pp 1–18

Bakshy E, Messing S, Adamic L (2015) Exposure to ideologically diverse news and opinion on Facebook. Sci 348(6239):1130–1132

Benlian A, Wiener M, Cram WA, Krasnova H, Maedche A, Möhlmann M, Recker J, Remus U (2022) Algorithmic management. Bus Inf Syst Eng 64(6):825–839

Berger B, Adam M, Rühr A, Benlian A (2021) Watch me improve – Algorithm aversion and demonstrating the ability to learn. Bus Inf Syst Eng 63(1):55–68

Biddle S, Ribeiro PV, Dias T (2020) Invisible censorship. Tiktok told moderators to suppress posts by "ugly" people and the poor to attract new users. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/. Accessed 8 Jan 2024

Bishop S (2021) Influencer management tools: Algorithmic cultures, brand safety, and bias. Social media+ society 7(1):1–13

Bloomberg (2022) Content moderation solutions market to cross US$ 32 Bn by 2031, TMR report. https://www.bloomberg.com/press-releases/2022-03-31/content-moderation-solutions-market-to-cross-us-32-bn-by-2031-tmr-report. Accessed 1 Nov 2022

Bogle A (2022) How Amazon has ended up funding far-right publishers and disinformation websites. https://www.abc.net.au/news/science/2022-09-15/amazon-affiliate-advertising-far-right-publishers/101436432. Accessed 2 Nov 2022

Bono JE, Judge TA (2003) Core self-evaluations: A review of the trait and its role in job satisfaction and job performance. Eur J Person 17(1):S5–S18

Bruns A (2021) Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. In: Hate speech and polarization in participatory society. Routledge, pp 33–48

Bucher T (2018) If then: algorithmic power and politics. Oxford University Press, Oxford

Cameron L, Lamers L, Leicht-Deobald U, Lutz C, Meijerink J, Möhlmann M (2023) Algorithmic management: Its implications

for information systems research. Commun Assoc Inf Syst 52(1):1–22
Chan TK, Cheung CM, Wong RY (2019) Cyberbullying on social networking sites: the crime opportunity and affordance perspectives. J Manag Inf Syst 36(2):574–609
Chen Y-S, Zaman T (2024) Shaping opinions in social networks with shadow banning. PLoS ONE 19(3):1–30
Chen A (2019) A leaked excerpt of Tiktok moderation rules shows how political content gets buried. https://www.technologyreview.com/2019/11/25/102440/tiktok-content-moderation-politics-protest-netzpolitik/. Accessed 29 Oct 2022
Columbres D (2023) Am I shadowbanned on Twitter? 3 quick solutions to fix it. https://tweetdelete.net/resources/am-i-shadowbanned-twitter-3-quick-solutions-to-fix-it/. Accessed 5 Jan 2023
Colve S (2018) Where did the concept of 'shadow banning' come from? https://www.vice.com/en/article/a3q744/where-did-shadow-banning-come-from-trump-republicans-shadowbanned. Accessed 29 Oct 2022
Cooban A (2023) EU stops advertising on X over hate speech. Fines could follow next year. https://amp-cnn-com.cdn.ampproject.org/c/s/amp.cnn.com/cnn/2023/11/22/tech/eu-advertising-x-hate-speech/index.html. Accessed 9 Jan 2024
Coombs WT (2007) Protecting organization reputations during a crisis: the development and application of situational crisis communication theory. Corp Reput Rev 10(3):163–176. https://doi.org/10.1057/palgrave.crr.1550049
Cotter K (2021) "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. Inf Commun Soc 26(6):1226–1243
Cram WA, Wiener M (2020) Technology-mediated control: case examples and research directions for the future of organizational control. Commun Assoc Inf Syst 46(1):70–91
Delmonaco D, Mayworm S, Thach H, Guberman J, Augusta A, Haimson OL (2024) "What Are You Doing, Tiktok?" How marginalized social media users perceive, theorize, and "prove" shadowbanning. In: Proceedings of the ACM on Human-Computer Interaction, pp 1–39. https://doi.org/10.1145/3637431
Douek E (2022) Content moderation as systems thinking. Harv Law Rev 136(2):526–607
Dreisbach T (2021) Alex Jones still sells supplements on Amazon despite bans from other platforms. https://www.npr.org/2021/03/24/979362593/alex-jones-still-sells-supplements-on-amazon-despite-bans-from-other-platforms. Accessed 22 Nov 2022
Duffy BE, Meisner C (2022) Platform governance at the margins: social media creators' experiences with algorithmic (in)visibility. Media Cult Soc 45(2):285–304
Elmimouni H, Skop Y, Abokhodair N, Rüller S, Aal K, Weibert A, Al-Dawood A, Wulf V, Tolmie P (2024) Shielding or silencing? An investigation into content moderation during the Sheikh Jarrah crisis. In: Proceedings of the ACM on Human-Computer Interaction, pp 1–21. https://doi.org/10.1145/3633071
Feuerriegel S, Dolata M, Schwabe G (2020) Fair AI. Bus Inf. Syst Eng 62(4):379–384
Fishman B (2023) Trust & safety as an enterprise: The decision spectrum and organizational structure. https://cinder.co/post/the-trust-safety-enterprise-decision-spectrum-functional-variability-and-organizational-structure. Accessed 8 Jan 2024
Fishman B, Harris LS (2023) Measuring trust & safety. https://www.cinder.co/blog-posts/measuring-trust-and-safety. Accessed 6 Jul 2024
Foucault M (2007) Security, territory, population: Lectures at the Collège De France, 1977–78. Springer, Cham
Franz A, Thatcher J. B (2023) Doxing and doxees: a qualitative analysis of victim experiences and responses. In: 31st European Conference on Information Systems (ECIS), Kristiansand, pp 1–16
Gillespie T (2022a) Do not recommend? Reduction as a form of content moderation. Soc Media Soc 8(3):1–13
Gillespie T (2022b) Reduction / borderline content / shadowbanning. Yale Law School Information Society Project, pp 1–14. https://law.yale.edu/sites/default/files/area/center/isp/documents/reduction_ispessayseries_jul2022.pdf
Gillespie T (2023) The fact of content moderation; or, let's not solve the platforms' problems for them. Media Commun 11(2):406–409
Gillespie T, Aufderheide P, Carmi E, Gerrard Y, Gorwa R, Matamoros-Fernández A, Roberts ST, Sinnreich A, West SM (2020) Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. Internet Policy Rev 9(4):1–30
Goldman E (2021) Content moderation remedies. Michigan Technol Law Rev 28(1):1–61
Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc 7(1):1–15
Gorwa R (2022) Stakeholders. Yale Law School Information Society Project. https://law.yale.edu/sites/default/files/area/center/isp/documents/stakeholders_ispessayseries_aug2022.pdf. Accessed 4 Nov 2022
Grimmelmann J (2015) The virtues of moderation. Yale J Law Technol 17:42–109
Guess AM, Malhotra N, Pan J, Barberá P, Allcott H, Brown T, Crespo-Tenorio A, Dimmery D, Freelon D, Gentzkow M, González-Bailón S, Kennedy E, Kim YM, Lazer D, Moehler D, Nyhan B, Rivera CV, Settle J, Thomas DR, Thorson E, Tromble R, Wilkins A, Wojcieszak M, Xiong B, de Jonge CK, Franco A, Mason W, Stroud NJ, Tucker JA (2023) Reshares on social media amplify political news but do not detectably affect beliefs or opinions. Sci 381(6656):404–408
Halckenhaeusser A, Foerderer J, Heinzl A (2020) Platform governance mechanisms: An integrated literature review and research directions. In: Proceedings of the 28th European Conference on Information Systems (ECIS), Online, pp 1–29
He Q, Hong Y, Raghu TS (2024) Platform governance with algorithm-based content moderation: An empirical study on Reddit. Inf Syst Res (In Print), pp 1–39
Hern A (2019) Revealed: How Tiktok censors videos that do not please Beijing. https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing. Accessed 5 Jan 2023
Hirsch F, Alizadeh A, Wiener M, Cram AW (2023) Algorithmic control in platform and traditional work settings: An updated conceptual framework. In: 31st European Conference on Information Systems (ECIS 2023), Kristiansand, pp 1–17
Jaidka K, Mukerjee S, Lelkes Y (2023) Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse. J Commun 73(2):163–178
Jaidka K, Mukerjee S, Lelkes Y (2021) An audit of Twitter's shadowban sanctions in the United States. In: 7th International Conference on Computational Social Science IC2S2, Zurich, pp 1–4
Jhaver S, Birman I, Gilbert E, Bruckman A (2019a) Human-machine collaboration for content regulation: the case of Reddit Automoderator. ACM Trans Comput-Hum Interact 26(5):1–35
Jhaver S, Bruckman A, Gilbert E (2019) Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. Proc ACM Hum-Comput Interact 3(CSCW). https://doi.org/10.1145/3359252
Jiang JA, Nie P, Brubaker JR, Fiesler C (2023) A trade-off-centered framework of content moderation. ACM Trans Comput-Hum Interact 30(1):1–34

Jorgenson D (2022) The Washington Post tried to get suppressed on Tiktok. Here's what happened. https://www.washingtonpost.com/technology/2022/10/28/tiktok-suppression/. Accessed 30 Oct 2022

Judge TA, Erez A, Bono JE, Thoresen CJ (2003) The core self-evaluations scale: development of a measure. Person Psychol 56(2):303–331

Katzenbach C (2021) "AI will fix this" – the technical, discursive, and political turn to AI in governing communication. Big Data Soc 8(2):1–8

Kaye DA (2019) Speech police: the global struggle to govern the internet. Columbia Global Reports, New York

Keller D (2019) Three constitutional thickets: why regulating online violent extremism is hard. George Washington University, Washington

Kellogg KC, Valentine MA, Christin A (2020) Algorithms at work: the new contested terrain of control. Acad Manag Ann 14(1):366–410

Kim A, Dennis AR (2019) Says who? The effects of presentation format and source rating on fake news in social media. MIS Q 43(3):1025–1039

Kim A, Moravec PL, Dennis AR (2019) Combating fake news on social media with source ratings: the effects of user and expert reputation ratings. J Manag Inf Syst 36(3):931–968

Kim TW, Routledge BR (2018) Informational privacy, a right to explanation, and interpretable AI. In: 2018 IEEE symposium on privacy-aware computing (PAC). IEEE, pp 64–74. https://doi.org/10.1109/PAC.2018.00013

Kizilcec RF (2016) How much information? Effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI conference on human factors in computing systems, pp 2390–2395. https://doi.org/10.1145/2858036.2858402

Klar R (2023) Musk's X will now allow verified users to hide check marks. https://thehill.com/policy/technology/4133090-musks-x-will-now-allow-verified-users-to-hide-checkmarks/. Accessed 9 Jan 2024

Lee MK (2018) Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data Soc 5(1):1–16

Leong C, Pan SL, Bahri S, Fauzi A (2019) Social media empowerment in social movements: power activation and power accrual in digital activism. Eur J Inf Syst 28(2):173–204

Lima C, Schaffer A (2022) Spotify has a white sumpremacist problem, Watchdog says. https://www.washingtonpost.com/politics/2022/09/23/spotify-has-white-supremacist-problem-watchdog-says/. Accessed 2 Nov 2022

Llewellyn C, Cram L, Hill RL, Favero A (2019) For whom the bell trolls: shifting troll behaviour in the Twitter Brexit debate. J Comm Mark Stud 57(5):1148–1164

Louis W, Thomas E, McGarty C, Lizzio-Wilson M, Amiot C, Moghaddam F (2020) The volatility of collective action: theoretical analysis and empirical data. Polit Psychol 41:35–74

Lutz S, Schneider FM (2021) Is receiving dislikes in social media still better than being ignored? The effects of ostracism and rejection on need threat and coping responses online. Media Psychol 24(6):741–765

Luu J (2023) This musician says his pro-Palestinian posts were banned. Is social media being censored? The Feed. https://www.sbs.com.au/news/the-feed/article/this-musician-says-he-was-shadowbanned-for-making-pro-palestinian-posts-is-social-media-being-censored/jipi1vvn3. Accessed 5 Jan 2023

Masnick M (2023) Elon Musk still loves 'shadow banning' those he doesn't like. https://www.techdirt.com/2023/03/17/elon-musk-still-loves-shadow-banning-those-he-doesnt-like/. Accessed 12 Jul 2024

McDonald B (2022) Extremists are seeping back into the mainstream: Algorithmic detection and evasion tactics on social media platforms. https://gnet-research.org/2022/10/31/extremists-are-seeping-back-into-the-mainstream-algorithmic-detection-and-evasion-tactics-on-social-media-platforms/. Accessed 2 Nov 2022

McKelvey F, Hunt R (2019) Discoverability: toward a definition of content discovery through platforms. Soc Media Soc 5(1):1–15

Merlan A (2020) How shadowbanning went from a conspiracy theory to a selling point. https://www.vice.com/en/article/v7gq4x/how-shadowbanning-went-from-a-conspiracy-theory-to-a-selling-point-v27n3. Accessed 29 Oct 2022

Merrer EL, Morgan B, Trédan G (2021) Setting the record straighter on shadow banning. In: IEEE Conference on Computer Communications, pp 1–10. https://doi.org/10.1109/INFOCOM42981.2021.948879

Milli S, Carroll M, Pandey S, Wang Y, Dragan AD (2023) Twitter's algorithm: Amplifying anger, animosity, and affective polarization. arXiv preprint, arXiv:2305.16941

Myers West S (2018) Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. New Media Soc 20(11):4366–4383

Möhlmann M (2021) Algorithmic nudges don't have to be unethical. Harv Bus Rev 22:1–7

Möhlmann M, Zalmanson L, Henfridsson O, Gregory RW (2021) Algorithmic management of work on online labor platforms: When matching meets control. MIS Q 45(4):1999–2022

Möhlmannn M, de Lima A, Salge C, Marabelli M (2023) Algorithm sensemaking: how platform workers make sense of algorithmic management. J Assoc Inf Syst 24(1):35–64

Newton C (2023) How Twitter keeps competitors off its For You Page. https://www.platformer.news/tiktok-nears-the-endgame/. Accessed 12 Jul 2024

Nicholas G (2022) Shedding light on shadowbanning. Center for Democracy & Technology (CDT), pp 1–52. https://doi.org/10.31219/osf.io/xcz2t. Accessed 29 Oct 2022

Nicholas G (2023) Sunsetting 'shadowbanning'. Yale Law School Information Society Project, pp 1–11. https://law.yale.edu/sites/default/files/area/center/isp/documents/sunsettingshadowbanning_ispessayseries_2023.pdf. Acccessed 20 Feb 2024

Oksanen A, Kaakinen M, Minkkinen J, Räsänen P, Enjolras B, Steen-Johnsen K (2020) Perceived societal fear and cyberhate after the November 2015 Paris terrorist attacks. Terror Polit Violence 32(5):1047–1066

Perez S (2023) Musk says X will address shadowbanning 'soon,' but former trust & safety exec explains why that will be difficult. https://techcrunch.com/2023/08/17/musk-says-x-will-address-shadowbanning-soon-but-former-trust-safety-exec-explains-why-that-will-be-difficult/. Accessed 5 Jan 2023

Riemer K, Peter S (2021) Algorithmic audiencing: why we need to rethink free speech on social media. J Inf Technol 36(4):409–426

Risius M, Blasiak KM, Wibisono S, Louis WR (2024) The digital augmentation of extremism: reviewing and guiding online extremism research from a sociotechnical perspective. Inf Syst J 34(3):931–963

Robertson RE, Green J, Ruck DJ, Ognyanova K, Wilson C, Lazer D (2023) Users choose to engage with more partisan news than they are exposed to on Google search. Nat 618(7964):342–348

Russo G, Ribeiro MH, Casiraghi G, Verginer L (2023) Understanding online migration decisions following the banning of radical communities. In: Proceedings of the 15th ACM Web Science Conference 2023 (WebSci 2023), Austin, pp 251–259

Ryan RM, Deci EL (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 55(1):68–78

Ryan-Mosley T (2023) Meta is giving researchers more access to Facebook and Instagram data. https://www.technologyreview.com/2023/11/21/1083760/meta-transparency-research-database-nick-clegg/. Accessed 9 Jan 2024

Ryan F, Fritz A, Impiombato D (2020) Tiktok and Wechat: Curating and controlling global information flows. 37/2020, Australian Strategic Policy Institute (ASPI), pp 1–72. https://ad-aspi.s3.ap-southeast-2.amazonaws.com/2020-09/TikTok%20and%20WeChat.pdf?VersionId=7BNJWaoHImPVE.6KKcBP1JRD5fRnAVTZ. Accessed 29 Oct 2022

SOMA (2024) U.S. 2020 Facebook and Instagram election study. https://somar.infoready4.com/#freeformCompetitionDetail/1910437. Accessed 9 Jan 2024

Sarker S, Chatterjee S, Xiao X, Elbanna A (2019) The sociotechnical axis of cohesion for the is discipline: its historical legacy and its continued relevance. MIS Q 43(3):695–720

Savolainen L (2022) The shadow banning controversy: perceived governance and algorithmic folklore. Media Cult Soc 44(6):1091–1109

Scrivens R, Gaudette T (2024) Online terrorism and violent extremism. In: Scrivens R, Gaudette T (eds) Oxford research encyclopedia of criminology and criminal justice. Oxford University Press, Oxford

Seering J (2020) Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. Proc ACM Hum-Comput Interact 4(CSCW2):1–28. https://doi.org/10.1145/3415178

Siapera E, Viejo-Otero P (2021) Governing hate: Facebook and digital racism. Telev New Media 22(2):112–130

Silva C (2022) Is your post blocked from being recommended? Insta Will Let You Know. Thanks... I guess? https://mashable.com/article/instagram-recommended-posts-blocked. Accessed 5 Jan 2024

Spiekermann S, Krasnova H, Hinz O, Baumann A, Benlian A, Gimpel H, Heimbach I, Köster A, Maedche A, Niehaves B, Risius M, Trenz M (2022) Values and ethics in information systems. Bus Inf Syst Eng 64(2):247–264

Stack L (2018) What is a 'shadow ban', and is Twitter doing it to Republican accounts. https://www.nytimes.com/2018/07/26/us/politics/twitter-shadowbanning.html. Accessed 29 Oct 2022

Starbird K, Arif A, Wilson T (2019) Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. Proc ACM Hum-Comput Interact 3(CSCW):1–26. https://doi.org/10.1145/3359229

Sullivan R, Veen A, Riemer K (2024) Furthering engaged algorithmic management research: surfacing foundational positions through a hermeneutic literature analysis. Inf Organ 34(4):1–20

Sunstein CR (2018) #Republic: divided democracy in the age of social media. Princeton University Press, Princeton

Suzor NP, West S, Quodling A, York J (2019) What do we mean when we talk about transparency? Towards meaningful transparency in commercial content moderation. Int J Commun 13:1526–1543

TikTok (2024) Research API. TikTok for developers. https://developers.tiktok.com/products/research-api/. Accessed 9 Jan 2024

Uber (2023) Understanding your restaurant's visibility. https://www.uber.com/en-AU/blog/understanding-your-restaurants-visibility/. Accessed 18 Sep 2024

Walsh A (2022) Tiktok censoring LGBTQ, Nazi terms in Germany: Report. https://www.dw.com/en/tiktok-censoring-lgbtq-nazi-terms-in-germany-report/a-61237610. Accessed 30 Oct 2022

Wankhede A (2022) Content moderation solutions market worth will reach US$ 26 Bn by 2031. https://www.linkedin.com/pulse/content-moderation-solutions-market-worth-reach-us-26-aditya-wankhede/. Accessed 1 Nov 2022

Wiener M, Cram WA, Benlian A (2023) Algorithmic control and gig workers: a legitimacy perspective of Uber drivers. Eur J Inf Syst 32(3):485–507

Wright L (2022) Automated platform governance through visibility and scale: on the transformational power of automoderator. Soc Media Soc 8(1):1–11

Zakharchenko K (2024) Facebook tries to combat Russian disinformation in Ukraine – FB Public Policy Manager. https://www.kyivpost.com/post/32048?ref=everythinginmoderation.co. Accessed 12 Jul 2024

Zannettou S (2021) "I won the election!". An empirical analysis of soft moderation interventions on Twitter. In: Proceedings of the 15th International AAAI Conference on Web and Social Media, pp 865–876. https://doi.org/10.1609/icwsm.v15i1.18110

Zeng J, Kaye DBV (2022) From content moderation to visibility moderation: a case study of platform governance on Tiktok. Policy Internet 14(1):79–95

Zuckerberg M (2021) A blueprint for content governance and enforcement. Facebook Notes. https://www.facebook.com/notes/751449002072082/. Accessed 12 Jul 2021