Interpreting Algorithmic Information Cues: User Sensemaking of Search Autocomplete Moderation

Shagun Jhaver[1]

Abstract

Autocomplete is a search feature that algorithmically generates information cues for any keywords entered in the search bar. While this feature makes the search process more efficient, it also frequently produces biased, misleading, offensive, or otherwise inappropriate suggestions. To address this problem, commercial search systems like Google Search now moderate Autocomplete information cues. However, we know relatively little about how users perceive this moderation process. Conducting interviews with 20 users of web search systems, I examine user attitudes toward the ethical tradeoffs in enacting moderation, reliance on search systems for making regulation decisions, and users' own role in moderating autocomplete information. My findings show that users desire greater visibility into the autocomplete moderation process. They also see flags and personal moderation mechanisms as promising avenues for themselves to exert greater agency within contemporary information infrastructures. My analysis bridges the fields of content moderation and search engine critiques, and lays the groundwork for enacting fair, accountable, and transparent search moderation.

[1]Rutgers University, Department of Library and Information Science, 184 College Ave, New Brunswick, NJ 08901. Phone: 972-400-1096.

Email: shagun.jhaver@rutgers.edu

Introduction

Search systems play a central role in shaping how people satisfy their information needs in everyday life (Cole 2011). Beyond returning ranked results, almost all mainstream search interfaces surface query autocomplete suggestions (Berget and Sandnes 2016) that structure users' sensemaking before any documents are accessed. These algorithmically generated cues function as informational signals that influence perceptions of relevance, credibility, and legitimacy (Markham 2024; Miller and Record 2017), yet how users interpret the control practices that shape their quality remains poorly understood. As search autocompletes become ever more widely deployed, understanding how users perceive their opaque production process is critical for theories of information behavior, trust, and accountability in search infrastructures.

While users highly trust search outputs, autocompletes can include inappropriate suggestions, such as misinformation, biases against specific groups, and stereotypes (Baker and Potts 2013; Leidinger and Rogers 2023; Lin et al. 2023). This is concerning because prior research has shown that across several writing tasks, exposure to biases in automated text suggestions tends to shift people's attitudes in the direction of the bias (Arnold et al. 2018; Jakesch et al. 2023; Williams-Ceci et al. 2024). Such shifts in attitude occur irrespective of users' awareness of the bias (Williams-Ceci et al. 2024). Thus, there is a need to "moderate", i.e., prevent inappropriate autocompletes from appearing before the users. In fact, search systems *do* moderate Autocomplete outputs, as I describe in the next section, yet the complexities of these moderation procedures usually remain hidden.

The prospect of moderating Autocomplete suggestions raises a range of ethical, technical, and political questions, such as how to distinguish between appropriate and inappropriate suggestions, who should have the power to make such distinctions, and how these decisions should be communicated to end-users. This paper takes a user-centered approach to interrogating these questions. By conducting semi-structured interviews with 20 regular users of commercial web search systems, I examine how users make sense of Autocomplete moderation, what concerns they have about its procedures, and how they seek to assert greater agency within the information retrieval process. This study contributes to information science by highlighting the perceived role of reporting mechanisms in shaping Autocomplete cues, capturing reflections regarding user control over the retrieved information, and exploring the distribution of curation responsibility (alongside the associated digital labor) across stakeholders. It advances understanding of users' transparency and accountability concerns about search moderation, and offers design and governance recommendations for how search systems can foster greater user trust.

Controlling Algorithmic Information Cues in Search Systems

General-purpose web search systems are essential mediators of information online. The frequent use of these systems can reinforce prevailing norms and narratives, showing them in a way that feels familiar and trustworthy to users. Unfortunately, a large body of prior research has documented that these narratives are prone to inaccuracies and biases (Graham 2022; Kay et al. 2015; Noble 2018). Seeing biased search results over and over can distort how people view the world (Noble 2018). Crucially, evidence suggests that exposure to identity-based stereotypes in search outputs can lead users to behave in ways that reinforce societal inequalities (Roy and Ayalon 2020).

Prior examinations of the outputs of Autocomplete, the feature I focus on for this article, have also reported derogatory suggestions for queries about gender identity, sexual orientation, race, women, old men, and religions (Baker and Potts 2013; Leidinger and Rogers 2023; Lin et al. 2023). Failure to identify and remove such information cues can harm not only searchers who develop skewed beliefs, but also the target(s) of the query, whose dignity and autonomy may be compromised by suggestions spreading false information about them (Miller and Record 2017). On the other hand, excessive or misused regulation could amount to censorship, as seen in authoritarian regimes that tightly control local search engines (Makhortykh et al. 2022). This emphasizes the need for conducting fair and efficient moderation of search outputs, including autocomplete suggestions.

Research on content moderation has traditionally focused on social media sites (Jhaver et al. 2023; Morrow et al. 2022; West 2018), but scholarly interest in search moderation has recently been growing. While the regular content curation and ranking practices of search engines would always reduce (or amplify) the visibility of some search outputs over others, search moderation is concerned with specific algorithmic configurations that regulate (i.e., remove, downgrade, or warn against) certain search outputs due to their inappropriate (e.g., offensive, misleading) content (Urman et al. 2024). Search engines employ information moderation in an even more opaque fashion than social media sites (Gorwa et al. 2020). Yet, previous content and policy analyses have highlighted that search systems regularly remove problematic content, inform users about potentially dangerous websites, downgrade inappropriate outputs, and place warning banners at the top of questionable search results (Robertson et al. 2025; Urman et al. 2024).

Search systems moderate not just search results but also Autocomplete suggestions. Up until 2016, Google enacted emergency takedowns or patches of inappropriate information cues and offered public explanations of its actions, usually in response to press reports (Gibbs 2016). In 2017, Google implemented a 'Direct feedback' or 'flag' feature, that allows end-users to report autocompletes under one of the categories: "hateful," "sexually explicit," "violent or includes dangerous and harmful activity," or "other" (Gomes 2017).[1] Over the following years, Google also developed an Autocomplete moderation policy that, in its current form, prevents predictions violating its overall content policies (which

block content that is dangerous, harassing, hateful, vulgar, etc.), and additionally omits suggestions that are election-related, health-related, disparage named individuals, or make unsubstantiated accusations against any groups (Google 2025). I contrast these search engine-directed framings of 'inappropriate' suggestions with what regular users deem should be moderated in Autocomplete outputs.

Recent theoretical analyses have highlighted the challenges of moderating search autocompletes. For example, Graham (2023) raised a range of ethical questions regarding legally permissible suggestions, including whether search systems have a duty to regulate group stereotypes and aggregated patterns of discrimination, and whether users must be allowed to influence autocomplete suggestions. I examine how regular users consider these ethical dimensions of autocomplete moderation. Hazen et al. (2022) have described the social and technical challenges of moderating autocomplete cues at scale. They especially note the difficulty of setting the boundary between problematic and non-problematic cases, and ask whether suppressing inappropriate suggestions is a form of censorship. I explore regular users' perspectives on these challenges in my analysis. Another relevant strand of search engine critique—although not specifically focused on search autocompletes so far—has considered how the prevalent 'service-for-profit' business model and users' information dependencies serve the "capital accumulation cycle" (Fuchs 2011) of search platforms (Mager 2012). I assess how users perceive platforms' capitalist spirit (Boltanski and Chiapello 2005) in the context of using search autocompletes.

Previous empirical research on autocomplete moderation has focused on documenting the extent of search moderation across different content categories, largely relying on algorithmic audits. For example, Leidinger and Rogers (2023) used stereotype-eliciting queries to compare moderation of societal biases across three search systems—Google, Yahoo!, and DuckDuckGo. They found that age- and gender-related stereotypes remain under-moderated in autocompletes across all three sites. Liu et al. (2024) compared the moderation practices on Google and Baidu by collecting over 2,000 autocompletes for 146 unique social groups. They concluded that both search systems under-moderate negative stereotypes across a wide range of social categories.

These prior empirical efforts help establish the current state of what autocomplete moderation achieves in practice. However, we know relatively little about how regular users, i.e., the intended audience of these moderation attempts, perceive the procedures and outcomes of autocomplete moderation, and how they evaluate the ethical questions that such moderation raises. The current article seeks to fill this crucial gap.

User Perspectives on Information Moderation

Prior research on user perspectives of search moderation is scarce. One notable exception is a recent article by Urman et al. (2024) that surveyed users to analyze

the relationships between user characteristics (e.g., demographic markers, political leaning) and support levels for search moderation. The current article builds upon that research by qualitatively exploring users' understanding of and complex concerns about search moderation. While Urman et al. (2024) focus on moderation of search results, my investigation centers moderation of Autocomplete outputs. Following Urman et al. (2024), I draw from user-centered research on social media moderation to shape my inquiry, as I detail below.

Discussions of user perspectives often include questions about who bears the responsibility for information moderation. For example, Riedl et al. (2021a) conducted a survey to examine user attitudes about who should intervene against problematic comments on the social media pages of online news outlets. They found that users largely attribute this responsibility to social media sites and news organizations, rather than law enforcement or themselves. In another survey study, Jang et al. (2024) showed that users with anti-establishment beliefs are more likely to hold individual users, rather than the government or platforms, responsible for social media content. Building upon these studies, I examine how users perceive the responsibilities of various stakeholders, including search systems, website providers, lawmakers, and users, regarding autocomplete moderation.

User attitudes toward content moderation are often shaped by third-person effects, i.e., a perception that media messages will be be more impactful or persuasive to others than to oneself (Davison 1983). Multiple studies have shown that presumed effects of inappropriate social media content on others significantly predict users' support for information moderation (Jhaver and Zhang 2023; Jhaver 2025a; Riedl et al. 2021b). I explore how concerns about the impact of search autocompletes on others shape users' attitudes toward autocomplete moderation.

Enacting content moderation involves making trade-offs between two fundamental values: upholding free speech and preventing harm caused by exposure to inappropriate content. Using a survey experiment, Kozyreva et al. (2023) showed that a majority of users prefer removing social media posts containing harmful misinformation over protecting free speech. On the other hand, support for free speech does not necessarily indicate aversion to content moderation (Guo and Johnson 2020; Jhaver 2025a). In the context of search autocompletes, while users' own speech is not at stake, they may still desire to preserve access to relevant autocompletes during search even if they represent norm-violating information cues. Therefore, I analyze how end-users conceive the dilemma of protecting free speech while preventing harms in autocomplete moderation.

Search platforms provide users an option to flag or report inappropriate autocompletes. However, flags are not unique to search. Indeed, all popular digital platforms offer flags as a feature that lets users directly request site administrators to take down any content (Zhang et al. 2023). The availability of flags introduces a novel 'rights-obligations' tension in content moderation, where users feel they have

a democratic right to flag but also experience flagging as a moral burden (Zhang et al. 2023).

Prior research has shown that users are motivated to flag inappropriate content due to an inclination to nurture their communities and a belief in generalized reciprocity (Zhang et al. 2023). Yet, users have reservations about the cognitive burden of flagging and the misuse of flags by bad actors attempting to remove otherwise appropriate content. Flagging interfaces require users to select one of the pre-defined categories of policy violations (e.g, "hate speech," "violent content") sustained by the flagged item. However, Kou and Gui (2021) found that the conception of what makes an item "flaggable" for users may differ from platforms' definitions of what is inappropriate. Chipidza and Yan (2022) show that flagging inappropriate content posted by prominent individuals might be even counterproductive in curbing its spread. Investigations of procedural fairness in flagging mechanisms show that users feel frustrated with platforms' lack of feedback after reporting, need justifications for flag outcomes, and desire greater transparency in each stage of the flag review process (Shim and Jhaver 2026; Zhang et al. 2023). I build upon this research to study users' expectations of flagging procedures regarding search autocompletes.

On social media sites, flags are often accompanied by a variety of personal moderation tools that allow users to "configure content moderation of the posts they see to align with their content preferences" (Jhaver and Zhang 2023). This includes tools that let users mute configured keywords (Jhaver et al. 2022), block offensive accounts (Geiger 2016), remove NSFW (not safe for work) content, and indicate their sensitivity to specific topical categories, such as hate speech and sexually explicit content (Jhaver et al. 2023). By empowering users to make their own moderation choices, these tools respond to a frequent platform critique that a centralized, one-size-fits-all approach to shaping content cannot serve the varied moderation needs of all users (Jiang et al. 2021).

Studies on user interactions with personal moderation tools have found that they provide users greater control over their social media feeds (Jhaver et al. 2018, 2022), yet fear of missing out on relevant posts makes users wary of using them (Jhaver et al. 2023; Jhaver 2025c). Jhaver and Zhang (2023) showed that both third-person effects and free speech support predict support for using personal moderation to regulate inappropriate content. While search platforms do not currently provide any personal moderation tools, I evaluate users' appetite for the customization that such tools could enable in moderating autocompletes.

Methods

This article is part of a larger project examining regular search engine users' understanding of how autocompletes are (1) produced and (2) moderated. In an earlier article (Jhaver 2025b), I have described users' mental models of how autocompletes are *produced*, and how these mental models raise critical concerns about user privacy, data manipulation, and reproduction of societal biases. During

my analysis, it became apparent that a separate group of findings, centered around users' understanding of how autocompletes are *moderated*, need to be explored in their own right. The current article documents these latter findings.

The data for this project are drawn from 20 semi-structured interviews conducted with regular users of commercial search systems. I recruited interview participants using a pre-screening survey that was circulated both online and through word-of-mouth, and received 161 responses. This survey asked for respondents' demographic details, experience with using search platforms, awareness of search autocompletes, and proficiency in information fields (i.e., whether they have worked or been educated in information technology (IT)-related fields). In line with prior research on this topic (Juneja et al. 2024), this study sought to understand the viewpoints of users who are not IT-experts. During recruitment, I drew on a theoretical sampling approach (Wengraf 2001) to guide my participant selection. This involved examining emerging insights and identifying gaps, attending to survey respondents' experiences and perceptions of Autocomplete feature, and seeking variation rather than representativeness. All participants resided in the US at the time of the interview. Table 1 (in Supplement) shows the demographic details of interview participants.

Data collection occurred between March–July, 2025. Each interview began with asking participants which search systems they use. Everyone (except P9) listed Google Search as their most frequently used search engine. Next, I asked interviewees to list the factors that they believe influence the production of autocomplete suggestions. My findings related to the discussions about these factors are captured in a separate article (Jhaver 2025b). In summary, participants listed three main factors: (1) searcher's personal search history and profile (including geographic location), (2) aggregate population-wide queries, and (3) commercial advertising.

Following this discussion, I asked participants questions (detailed below) related to the moderation of inappropriate search autocompletes for the rest of the interview. Participants' responses to these moderation-related questions constitute the findings for the current article. Given the widespread assumption that population-wide queries influence autocomplete production, my questions about autocomplete moderation began with the following prompt:

> *Would you prefer to see suggested queries submitted by other users even when they are biased, offensive, untrue, or problematic in other ways, or would you prefer search platforms to remove such problematic suggestions?* ***Explain why.***

Next, to contextualize discussions about autocomplete moderation, I showed participants a few illustrations of problematic autocompletes (e.g., suggestions displaying negative stereotypes of Europeans, Democrats, and transgender people) that have been documented in prior research (Olteanu et al. 2020). I questioned participants who they blamed for such autocompletes appearing in search, and whether (and how) platforms should review and remove such information cues.

Following this, I asked questions about flagging (or reporting) as a defense mechanism against problematic suggestions: should platforms offer flags, whether all users have a duty to report inappropriate autocompletes, how they believe flag review process works, and what they expect of post-flagging outcomes. Finally, I inquired participants about the design and policy solutions search systems can enact to increase their trust in autocomplete suggestions. At the conclusion of each interview, participants were compensated for their time with a $20 Amazon gift card.

All interview transcripts were recorded, transcribed, cleaned, de-identified, and then uploaded to Dedoose, a qualitative analysis software. Once it became clear that I needed to investigate my findings about autocomplete moderation separately, I copied my Dedoose project, deleted existing codes, and restarted the coding process to prepare this article. I began with identifying relevant excerpts—those pertaining to search moderation—and then performed a reflexive thematic analysis on them using an inductive approach (Braun and Clarke 2006).

I adopted an experiential orientation to data interpretation and prioritized semantic (rather than latent) codes (Clarke and Braun 2014). Through an iterative coding process, memo-writing, and continual comparisons of codes with one another and with interview data, I allowed the codes to emerge and refine organically. My coding was informed by a constructionist epistemology: while I considered recurrence, I primarily based code development and application on participants' expressed significance of the issues discussed, alongside my interpretation of their meaningfulness and a reflexive analysis of my positionality. I did not seek inter-rater reliability with a second coder because my codes were merely an interim product in an iterative process that evolved over multiple cycles, and not a final result requiring testing (McDonald et al. 2019). Next, I refined and consolidated my codes, which resulted in four key themes. I engaged in negative case analysis, i.e., I looked for data that contradicted my themes and revised my themes to account for them. I conducted member checking to further increase the validity of my findings, i.e., I shared my manuscript drafts with participants, and confirmed that my interpretations resonated with their views and experiences.

Findings

Perceived Necessity of Autocomplete Moderation

Many participants observed that their general perception of search systems, and especially Google, as fact-finders makes them consider Autocomplete suggestions, at least at a first glance, as factually correct. For example, P17 felt that most people today equate Googling to "looking for truth," and thus would expect any information that appears in Google autocompletes to be credible. Similarly, P13 shared that she tends to take autocompletes at face value:

> *"There are certain things about which we're trained to be critical — like the opinion page [of newspapers]. It feels often as if a search bar*

> *is a neutral technology, so it flies under the radar, like okay, well, this isn't the time when I need to have my critical thinking hat on."* – P13

Given this expectation, participants sensed a danger in Autocomplete affirming incorrect or biased ideas by including them in its suggestions. Many worried that autocompletes that show negative associations for vulnerable groups (e.g., racial, sexual, or gender minorities) can "subconsciously brainwash" (P8) searchers to categorize those groups negatively. For instance, P9 shared:

> "*There is affinity for people to get associated with negative views a lot more easily compared to positive views, so perpetuating negative opinions about any identity group [via autocompletes] will make those opinions stick.*" – P9

Some participants felt that children and digitally illiterate users are particularly vulnerable to the effects of such risks, and should therefore be protected. For these participants, offering such protections requires that Autocomplete production regulates the display of inappropriate information cues. For instance, P8 feared that an absence on such regulation would flood autocompletes with fake news and misinformation. P10 emphasized that autocompletes should especially exclude any biased or offensive keywords when her search query does not pertain to any overtly political or ideologically charged topics.

When asked which suggestions should be omitted from appearing in autocompletes, participants reflected that any information that is not "true" or "decent" should be moderated. Specifically, they argued that any instances of stereotypes, false information, illegal content, and hate speech should be regulated. For instance, P5 noted:

> "*If it's abusive or inappropriate, it shouldn't appear. For example, if a teenager [is] typing "teenagers are..." and then sees something like "horrible," "lazy," or "disrespectful," then in one way or the other, it might affect or trigger the emotion or mentality of the person typing, so Google should avoid such suggestions.*" – P5

P5 further mentioned that autocompletes "should be relevant to what you're typing," and not "unrelated or expected." Similarly, P13 felt that any autocompletes that suggest "irrelevant stuff that's not nice or that are hurtful or disturbing" should not be shown to the users. P3 concurred that autocompletes that derogate any demographic or interest groups, such as Europeans or transgender individuals, could be upsetting for searchers from that group and lower their self-esteem, and should therefore be removed:

> "*I think, literally almost everyone uses Google to search, and it's used for many reasons, so it's important that Google maintains the decency [in its outputs] for its users. It shouldn't be a platform that promotes hate, that promotes discrimination, or promotes crime.*" – P3

Despite recognizing the need to moderate autocompletes, many participants also observed the potential pitfalls of such regulation, as I discuss next.

Tradeoffs in Moderating Search Autocompletes

Almost all participants felt that there are certain types of information—notably, child pornography and other illegal content—that are "beyond the pale" (P13) or directly "cause harm" (P11) and should not be suggested by autocompletes. Thus, the need for a baseline level of autocomplete moderation was broadly recognized. However, many participants hesitated to moderate autocompletes that fall in a "gray area" (P13, P17) of what could be considered misinformation or offensive speech.

These hesitations often connected to the perceptions of how search engines produce Autocomplete suggestions. For instance, most participants assumed that autocompletes at least partially reproduce population-wide searches, and thus they offer a lens into understanding society. Given this assumption and a desire to learn what others are thinking, some participants saw autocomplete moderation as a distortion of popular search trends, and therefore, opposed it. For instance, P11 described herself as a "curious person" who appreciates knowing uncomfortable or problematic truths about about any search topic, and disapproved of receiving "over-sanitized" autocompletes. P13 framed this as a free speech issue, arguing that "in a healthy democracy, users should be able to find a range of opinions [on any topic] if they are looking for it on the internet." P2 considered himself as capable of distinguishing truth from subjective opinions when consuming autocompletes, and felt that he would not be swayed by biased autocompletes. Similarly, P19 objected to regulating autocompletes that stereotype any groups by observing:

> *"I think that if we are told that people don't believe others are evil, then that's disingenuous. People do believe that, and they live that, they walk that, they search for that and share that. Of course, it's wrong to pick a group based on their thoughts or identifications, and then demonize them. But that's what humans are doing. And unless we're clear and open about it, it's not going to improve. You can't just censor it into oblivion."* – P19

On the whole, participants wished that search moderation would achieve an appropriate balance between making relevant information available (i.e., not censoring it) and ascertaining information quality (e.g., ensuring public safety) within Autocomplete outputs. However, they also acknowledged the difficulty of attaining this balance. P7 admitted that everyone has different views on where search engines should draw the line for regulating any candidate suggestion, and even an individual's views may evolve over time. P1 considered it impossible for search engines like Google to "review every single piece of information that's put on their search platform" and determine whether autocompletes derived from

it deserve moderation. P1 further remarked that different countries (or legal jurisdictions) likely put pressure on search companies to moderate as per their (potentially conflicting) priorities. Thus, it could be challenging to enact consistent and broadly satisfying search moderation. P9 offered:

> *"If you want to implement this properly, there is going to be a lot of political debate of what exactly constitutes things worth removing. And let's say that you come up with the idea of 'these specific things are bad.' Now, how do you go about implementing that? I mean, there would be technological challenges, societal challenges, political challenges, and feasibility challenges."* – P9

While participants appreciated these varied challenges of moderating autocompletes, they were also keenly aware of search systems' central role in enacting such moderation, and expressed nuanced views on it as I describe next.

Reliance on Search Systems to Moderate Autocompletes

Many participants trusted search platforms to be relatively objective and unbiased when curating autocomplete suggestions. P17 commented on the aesthetics of the Google Search homepage, noting that "the simplicity of it and the lack of forward facing advertising makes me feel that I'm going to get a truthful result out of it." P10 felt that Google Search just shows "what is most popular without applying its own judgment" in its autocompletes, and thus considered them to be politically neutral. Participants frequently compared search moderation with social media moderation and expressed a much more favorable view of the former. For example, P1 estimated that search system outputs derive from websites of organizations and agencies as opposed to what regular social media users might say, and are therefore, more reliable.

On the other hand, participants felt uneasy about ceding the information authority of making search moderation calls to platforms alone. Many reported having confidence in commercial search systems like Google a decade ago, but losing trust in their outputs in recent years. They evoked the commercial motivations and market monopoly of these systems as the reasons for their distrust in their moderation practices. For example, P19 referred to search platforms as "very highly capitalized companies" whose "algorithms are financially motivated" and expressed skepticism about their search outputs.

While these participants did not fully trust search systems with moderation, they also could not point to any other entity whose decisions would be unimpeachable in the current political climate. For example, P9 desired to see search moderation being conducted or at least overseen by a "trusted third party whose decisions are going to be unbiased," but admitted that such a third party would be "very hard to find." Similarly, P1 said:

> *"It would be great if there was some kind of like a pure organization that you could hundred percent trust to be neutral, to be intelligent, and*

> *to be thoughtful and really thorough. But I don't think that something like that exists quite yet. So I don't totally trust Google. But I also don't know who else I would trust to do this."* – P1

When considering responsibility allocation for inappropriate suggestions, most participants recognized that search systems hold the most power in shaping autocompletes, and thus held them responsible. For example, P6 blamed Google Search for biased Google autocompletes "because they're ultimately the ones that are relaying that information." Given this perceived responsibility, P17 desired search systems to publicly profess their commitment to actively reduce hate speech and misinformation in autocomplete suggestions. On this point, P2 presumed that search systems strategically avoid taking an overt public stance on or responsibility for moderation:

> "*I don't think they [search systems] should be so trusted to curate everything and moderate everything, which I think is consistent with the view of these companies themselves, right? They don't want to tread into that territory. I think it would, you know, cause them more trouble for them to accept that liability.*" – P2

Some participants noted that policymakers can play a crucial role as a counterbalancing agent by instituting minimum requirements on search systems, such as regulating extremist content and enacting greater transparency in autocomplete moderation. However, many worried about government overreach, arguing that policymakers themselves may be biased. For instance, P18 feared:

> "*I'm afraid the lawmakers we have now would not be trustworthy anyway, so they would just look at it as a way to take advantage and get the autocompletes they want.*" – P18

While platforms were often blamed for inappropriate autocompletes, many participants argued that users also have a responsibility to recognize that search outputs may contain misinformation or biases, and thus, they must do their due diligence (e.g., verifying sources, consulting different viewpoints) before forming beliefs, especially on ideologically charged topics. For example, P11 said:

> "*I think, in general, like on and off the Internet, we are responsible for what we believe in and I think it is really important, whether it's an autocomplete or a search result, that we take everything with, you know, some critical thinking.*" – P11

A few participants felt that search systems should not be held responsible for fact-checking information shown in autocompletes because they did not create this information themselves, but rather rely on external sources to populate Autocomplete suggestions. Thus, as per these participants, it is

the responsibility of these underlying sources—such as news websites—to provide reliable information. However, the absence of source data accompanying autocomplete suggestions made it difficult for participants to evaluate their reliability or ascribe responsibility for their inappropriateness to the sources used. To address this, participants desired search systems to be more transparent about how they curate and moderate autocompletes, and especially reveal which source each autocomplete derives from. For example, P17 reflected:

> "*I came from a time when we searched for books in libraries. We were taught to find the source material and then evaluate the validity of that source material. So based just on these autocompletes, I know I'm only getting a partial picture. I mean, it gives me some concern that I don't know the source. It may be a news source or, you know, some crazy guy's blog.*" – P17

Besides search systems, participants felt that end-users themselves could also be a significant stakeholder in autocomplete moderation, as I next describe.

Searchers' Role in Moderating Autocompletes

Participants discussed two key avenues in which users can play an active role in moderating or shaping search autocompletes: reporting (i.e., flagging) inappropriate suggestions and configuring controls that customize (i.e., personalize) suggestions. I describe each in turn below.

Reporting Autocompletes. Many participants desired to have a *flag* functionality that would allow them to report inappropriate autocomplete suggestions. They saw the value of this feature in empowering end-users to object and apprise search systems of norm-violating information cues. They hoped that search systems' actions on these reports could prevent other users from being exposed to similar suggestions. P9 supported the availability of reporting function, but characterized it as a "last stop-gap solution"; he wanted search systems to additionally institute systemic defensive mechanisms that proactively detect and remove inappropriate autocompletes.

None of my participants were aware that Google Search already provides an option to flag inappropriate autocompletes. Once they saw its current implementation, participants uniformly felt that the flag feature is inconspicuous by design and wanted Google to present it more prominently. When asked whether everyone has a *responsibility* to report norm violations they encounter, all participants (except P3) felt that it should be a right but not a duty for users to report inappropriate autocompletes. For example, P7 said:

> "*I don't think anybody should feel obligated to report. But if somebody feels strongly about it, they should have the ability to report stuff if they want to.*" – P7

In contrast to the above-mentioned supportive views of flags, some participants did not see any benefit in search systems offering flags, positing that very few users would ever use them, especially if flag submission is labor-intensive. Others felt that providing flags would actively cause harm. Notably, both P15 and P19 expressed similar concerns about flag abuse, observing that some users may report otherwise appropriate suggestions just because of their ideological objections, and possibly disrupt the neutrality of autocompletes.

Participants usually had low expectations of post-reporting outcomes based on their experiences with using flags on social media platforms. Many of them expected reports to be ignored and felt that an autocomplete will need to have multiple reports lodged against it to trigger any platform reviews. Some desired to receive regular updates about different stages of the flag review, review outcome, and an explanation of that outcome, but they did not expect search systems to offer such updates. P19 considered the flagging feature "a facade" that gives users "a sense of control that you don't actually have." Similarly, P17 shared:

> "*I think it makes you feel good in that moment. But in reality, we've all seen, you know, that two days later, where the platform is like, 'no, we didn't take this down,' even though it's, you know, obviously something that's terrible.*" – P17

Reflecting on flag reviews, participants appreciated search systems' challenge with scaling up the review process as the number of flag reports increases. They expected that search systems address this challenge by deploying a coordinated mix of human moderators, AI tools, and moderation policies to enact flag reviews. For example, P13 estimated:

> "*There should be some basic stuff that the AI can catch, but the more complicated political questions—and there are probably many of them that are much more complicated than what I'm even thinking about—those probably do require human review.*" – P13

Customizing Autocompletes. Again and again during the early interview sessions, my participants kept bringing up their need to customize search autocompletes, even though search systems do not currently offer such facility. Therefore, I extended my inquiry to understand this need and explore what this customization could look like within the search feature.

First, many participants objected to the fact that search systems do not allow turning off autocompletes altogether. For instance, P3 sometimes find autocompletes "just annoying" and wanted a search setting that could temporarily turn them off. Additionally, participants desired the ability to toggle on/off or customize their search outputs along a range of dimensions, including negativity, controversiality, toxicity, presence of pornographic content or swear words, and search location. These ideas mirrored the design of personal moderation tools available on social media platforms, such as the binary toggles for viewing sensitive

content on Twitter and 3-level (Show/Blur/Hide) controls for viewing violent and sexual content on Tumblr (Jhaver et al. 2023). For example, P7 pointed out:

> "*The user should be able to customize their search in any way that they want. I don't see a problem with just telling Google like, Hey, don't show me autocompletes based on my location. Don't show me autocompletes based on my own personal search history. Or maybe you know, you could even go so far as to say, like, only show me autocomplete suggestions based on XYZ.*" – P7

Many participants acknowledged that a majority of users tend to never change the default search settings, and they therefore considered it vital that search moderation ensures appropriate outputs even without any customization. Still, participants wanted the freedom to further shape search outputs for different use contexts. Participants' desire to customize often derived from their motivation to protect vulnerable groups, especially children. For example, P7, P11 and P12 saw search customization as a form of "parental control" that could prevent children from being exposed to inappropriate suggestions. P11 offered:

> "*I think if parents had more control over it like, you know, specific things that they wouldn't want their child to encounter on Google. I think that could be helpful.*" – P11

Interview data also revealed the potential drawbacks of customizing autocompletes. Some participants appreciated the plainness and simplicity of Google Search interface and feared that customization would make their search experience more complicated. Others felt concerned that enabling search customization would lead to information insularity. For example, P16 worried that this customization would lead people to see "only part of the story." Similarly, P17 said:

> "*I think that the hard part is that you're only going to find what you're looking for, and the downside of that is, it limits discovery. And so I think, doing that, you're not going to maybe find an enriching path that you would not have thought of.*" – P17

Discussion

My analysis shows a broad consensus for instituting at least a baseline level of search moderation to detect and remove overtly inappropriate autocompletes, e.g., those that lead to illegal content. Similar to social media users (Jiang et al. 2023; Kozyreva et al. 2022; Zhang et al. 2023), my participants relied on their ethical values (i.e., a desire for "truth" and "decency") and prior media use to conceive of additional content types (e.g., fake news, identity-based stereotypes) that should be regulated within search information cues. These user conceptions

of what moderation ought to achieve only partially aligns with search systems' moderation policies. For example, Google Search—in alignment with user needs—seeks to moderate hateful and vulgar suggestions, but it additionally regulates other content types, e.g., all election-related suggestions, which indicates its focus on avoiding litigation rather than serving end-users.

My participants recognized that suggestions that fall in the gray area of moderation pose the challenge of determining where (and how) to draw the lines for information removal. They viewed search moderation as a compromise between ascertaining information quality and having access to diverse viewpoints, yet they found it challenging to envisage how such compromise can be achieved or what steps it would entail. This relates to the tension between upholding free speech and preventing content-based harms that has been explored in prior social media moderation research (Guo and Johnson 2020; Jhaver 2025a; Kozyreva et al. 2023).

Relatedly, the question of agency looms large in my interview data: *who* exactly should have the authority and responsibility to draw the lines for autocomplete removals? On this question, in line with prior research (Riedl et al. 2021a), participants largely held platforms responsible for inappropriate autocompletes, and expected them to make sensible decisions, e.g., removing instances of hate speech and misinformation. Indeed, participants reported a higher expectation of search moderation as compared to social media moderation.

On the other hand, participants appreciated the political, technological, and scaling challenges that search systems face in enacting autocomplete moderation. They also realized the pivotal roles that other stakeholders, including website providers, lawmakers, and users themselves, must play to improve the quality of search outputs. This aligns with a "pluralist model of speech regulation" (Balkin 2017), and a growing recognition that moderation must now occur in a multi-stakeholder fashion (Jang et al. 2024; Riedl et al. 2021a). At the same time, it is important to recognize the power asymmetries between search engines and other stakeholders, especially users—it is search systems' responsibility to institute processes that allow and encourage other stakeholders to help regulate search outputs.

My interview data show that users are increasingly losing confidence in search moderation, especially in more recent years. While participants still put greater trust in search engines as compared to social media platforms, this growing mistrust may reflect a response to the capitalist logics (Mager 2012) of search systems that are getting increasingly visibilized through news reports and academic research. This is concerning, especially given the outsized dependency that almost everyone has today on commercial search systems as information intermediaries for almost every topic. My findings also suggest that users feel a sense of resignation regarding autocomplete moderation because they do not see any viable alternatives to the current reliance on search systems, neither do they view lawmakers as necessarily trustworthy. This predicts a difficult sociotechnical challenge in negotiating the balance between online safety and information access (Kozyreva et al. 2023).

My analysis extends prior research on how third-person effects shape users' moderation preferences (Lim et al. 2025; Riedl et al. 2021b). I found that concerns for vulnerable groups, especially children and digitally illiterate users, strengthened participants' support for autocomplete moderation and their motivation to report inappropriate autocompletes. Further, their desire for autocomplete customization also often derived from their need to enact parental controls in search outputs. Future work can build upon these findings by conducting a more focused exploration of how anxieties about children's information behaviors shape parents' and non-parents' search moderation preferences.

Participants deemed flagging mechanisms as a valuable regulation approach that can empower end-users to have a voice in search moderation. However, all participants agreed that the discoverability of flags is severely limited within popular search interfaces. This is clearly a conscious choice on the part of search systems, and it reflects their stark lack of transparency or investment in user education regarding search moderation. This design choice is perhaps an attempt by search systems to reduce the burden of flag reviews. Yet, it leaves harmed individuals and groups without a clear avenue to complain about problematic autocompletes. Further, in line with prior research (Bäumler et al. 2025), my participants admitted that the cognitive labor required to submit flags would be a key consideration in their decision to flag inappropriate suggestions. Participants also worried about the exploitation of flags by bad actors. Thus, search systems must design their autocomplete flags to be more noticeable and easy to use, and defend against false flagging.

Echoing prior findings on the user-needs of social media flags (Bäumler et al. 2025; Shim and Jhaver 2026; Zhang et al. 2023), my participants desired to receive regular updates about flag reviews and an explanation of review outcomes. Additionally, participants wanted more visibility into the autocomplete moderation process, including access to sources behind each suggestion. This indicates that the principles of procedural fairness, accountability, and transparency, which have so frequently guided recent social media moderation efforts (Schoenebeck et al. 2021; Shim and Jhaver 2026; Vaccaro et al. 2021) continue to be relevant for search moderation as well.

Besides procedural transparency, my participants also frequency evoked a need for greater control and customizability in autocomplete moderation. This became most apparent in their frequent demand for personal moderation tools to configure search autocompletes. In line with prior research (Heung et al. 2025; Jhaver et al. 2023; Jhaver 2025c), fear of missing out on relevant information and concerns about being placed within information bubbles deterred interviewees' interest in such tools. It would be valuable to investigate further how such concerns are shaped by searchers' information-seeking goals, and how a more granular customization could assuage such concerns. Similar to findings from prior research on social media content (Heung et al. 2025; Jhaver et al. 2023), my participants expected a baseline level of system review that would detect and remove blatantly

inappropriate autocompletes for all users, regardless of their personal moderation configurations. On the whole, my analysis suggests that introducing personal moderation within search systems could be a fruitful avenue for search engines to empower their users and foster trust.

From an information science perspective, my findings extend existing accounts of information behavior (Cole 2011; Huvila and Gorichanaz 2025) by highlighting how users engage in pre-retrieval sensemaking when interacting with algorithmic nudges within search systems. Rather than treating autocomplete suggestions as neutral conduits to information, users interpret them as opaquely derived system outputs that may carry institutional intent, conceive of situations when these outputs are inappropriate, and derive from their use of social media sites to envision how such outputs should be moderated. This study contributes to research on human–algorithm interaction and information retrieval (Chen and Tang 2025; Jiang et al. 2025) by showing how user imaginaries of algorithmic moderation influence their trust in search systems and shape expectations about their own role in shaping search infrastructures.

Limitations and Future Work

The recruitment channels and theoretical sampling approach used in this study shaped the composition of the participant sample. In particular, this sample is more highly educated than the general population and may therefore possess greater training in critical analysis and ethical reasoning. Moreover, this sample comprises only the users living in the US, and its societal values and media climate could have influenced my findings regarding users' perspectives on free speech, platform responsibilities, and Autocomplete customization. Note, however, that the goals of this study did not call for constructing a representative sample. Instead, I sought to develop a nuanced understanding of the key concepts and concerns in Autocomplete moderation rather than to achieve broad generalizability. Further research would be helpful to assess the extent to which the perspectives and mental models surfaced in my findings transfer to different countries and demographic groups.

Conclusion

This article has presented an empirical analysis of how users make sense of Autocomplete moderation, how their concerns for vulnerable others shape their attitudes, and how they view search systems' and their own responsibilities within this moderation process. The nuanced insights presented here offer first steps toward creating a blueprint for building more fair, accountable, and transparent search moderation.

Given the ubiquitous use of search engines and the emerging popularity of generative AI tools, I propose that Autocomplete moderation is a promising research site to examine the ethics of knowledge production within human-AI interactions. For example, it would be valuable to investigate whether (and how)

to regulate information cues that derive from emerging news stories, that are not politically neutral, or that present inter-regional inconsistencies due to algorithmic localization. More broadly, I call for studies that involve end-users in designing content moderation for systems that serve as conversation partners.

Declarations
No use of Generative AI occurred in any stage or process of conducting and writing this research.

Notes

1. These flagging categories have evolved multiple times since 2017.

References

Arnold, K. C., Chauncey, K., and Gajos, K. Z. (2018). Sentiment bias in predictive text recommendations results in biased writing. In *Graphics interface*, pages 42–49.

Baker, P. and Potts, A. (2013). 'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms. *Critical discourse studies*, 10(2):187–204.

Balkin, J. M. (2017). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UCDL Rev.*, 51:1149.

Bäumler, J., Bader, H., Kaufhold, M.-A., and Reuter, C. (2025). Towards youth-sensitive hateful content reporting: An inclusive focus group study in germany. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Berget, G. and Sandnes, F. E. (2016). Do autocomplete functions reduce the impact of dyslexia on information-searching behavior? The case of Google. *Journal of the Association for Information Science and Technology*, 67(10):2320–2328.

Boltanski, L. and Chiapello, E. (2005). The new spirit of capitalism. *International journal of politics, culture, and society*, 18(3):161–188.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101–77–101.

Chen, P.-Y. and Tang, M.-C. (2025). The Influence of Recommendation Strategies and Music Diversity on User Preference: A Structural Equation Modeling Approach to Subjective and Objective Measures. *Proceedings of the Association for Information Science and Technology*, 62(1):117–126.

Chipidza, W. and Yan, J. (2022). The effectiveness of flagging content belonging to prominent individuals: The case of Donald Trump on Twitter. *Journal of the Association for Information Science and Technology*, 73(11):1641–1658.

Clarke, V. and Braun, V. (2014). Thematic analysis. In Teo, T., editor, *Encyclopedia of critical psychology*, page 1947–1952. Springer New York.

Cole, C. (2011). A theory of information need for information retrieval that connects information to knowledge. *Journal of the American Society for Information Science and Technology*, 62(7):1216–1231.

Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly*, 47(1):1–15.

Fuchs, C. (2011). A contribution to the critique of the political economy of google. *Fast Capitalism*, 8(1).

Geiger, R. S. (2016). Bot-based collective blocklists in twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6):787–803.

Gibbs, S. (2016). Google alters search autocomplete to remove "are jews evil" suggestion. *The Guardian*, 5.

Gomes, B. (2017). Our latest quality improvements for search. *Google blog*, 25.

Google (2025). How google autocomplete predictions work.

Gorwa, R., Binns, R., and Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*, 7(1).

Graham, R. (2022). *Investigating Google's Search Engine: Ethics, Algorithms, and the Machines Built to Read Us*. Bloomsbury Publishing.

Graham, R. (2023). The ethical dimensions of google autocomplete. *Big Data & Society*, 10(1):20539517231156518.

Guo, L. and Johnson, B. G. (2020). Third-person effect and hate speech censorship on facebook. *Social Media + Society*, 6(2):2056305120923003.

Hazen, T. J., Olteanu, A., Kazai, G., Diaz, F., and Golebiewski, M. (2022). On the social and technical challenges of web search autosuggestion moderation. *First Monday*, 27(2).

Heung, S., Jiang, L., Azenkot, S., and Vashistha, A. (2025). "Ignorance is not Bliss": Designing Personalized Moderation to Address Ableist Hate on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Huvila, I. and Gorichanaz, T. (2025). Trends in information behavior research, 2016–2022: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 76(1):216–237.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. (2023). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15.

Jang, H., Barrett, B., and McGregor, S. C. (2024). Social media policy in two dimensions: understanding the role of anti-establishment beliefs and political ideology in americans' attribution of responsibility regarding online content. *Information, Communication & Society*, 27(6):1047–1072.

Jhaver, S. (2025a). Bans vs. warning labels: Examining bystanders' support for community-wide moderation interventions. *ACM Trans. Comput.-Hum. Interact.*,

32(2).

Jhaver, S. (2025b). Examining how search engine users understand the production of autocomplete suggestions. *New Media & Society*, 0(0):14614448251406282.

Jhaver, S. (2025c). Personal moderation configurations on facebook: Exploring the roles of fear of missing out, social media addiction, norms, and platform trust. *First Monday*.

Jhaver, S., Chen, Q. Z., Knauss, D., and Zhang, A. X. (2022). Designing word filter tools for creator-led comment moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1 – 21, New York, NY, USA. Association for Computing Machinery.

Jhaver, S., Ghoshal, S., Bruckman, A., and Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput.-Hum. Interact.*, 25(2). 1073-0516.

Jhaver, S. and Zhang, A. (2023). Do users want platform moderation or individual control? examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society*.

Jhaver, S., Zhang, A., Chen, Q. Z., Natarajan, N., Wang, R., and Zhang, A. (2023). Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Jiang, J. A., Nie, P., Brubaker, J. R., and Fiesler, C. (2023). A trade-off-centered framework of content moderation. *ACM Trans. Comput.-Hum. Interact.*, 30(1):Article 3.

Jiang, J. A., Scheuerman, M. K., Fiesler, C., and Brubaker, J. R. (2021). Understanding international perceptions of the severity of harmful content online. *PloS one*, 16(8):e0256762.

Jiang, T., Sun, Z., and Fu, S. (2025). Restraining the formation of filter bubbles with algorithmic affordances: Toward more balanced information consumption and decreased attitude extremity. *Journal of the Association for Information Science and Technology*.

Juneja, P., Zhang, W., Smith-Renner, A. M., Lamba, H., Tetreault, J., and Jaimes, A. (2024). Dissecting users' needs for search result explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 3819–3828.

Kou, Y. and Gui, X. (2021). Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community.

Kozyreva, A., Herzog, S., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., and Reifler, J. (2022). Free speech vs. harmful misinformation: Moral dilemmas in online content moderation. *Proceedings of the National Academy of Sciences of the United States of America*.

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., and Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120.

Leidinger, A. and Rogers, R. (2023). Which stereotypes are moderated and under-moderated in search engine autocompletion? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1049–1061.

Lim, J. S., Lee, C., Kim, J., and Zhang, J. (2025). Influence of covid-19 vaccine misinformation beliefs on the third-person effect: implications for social media content moderation and corrective action. *Online Information Review*, 49(3):497–516.

Lin, C., Gao, Y., Ta, N., Li, K., and Fu, H. (2023). Trapped in the search box: An examination of algorithmic bias in search engine autocomplete predictions. *Telematics and Informatics*, 85:102068.

Liu, G., Pinoli, P., Ceri, S., and Pierri, F. (2024). A comparison of online search engine autocompletion in google and baidu. *arXiv preprint arXiv:2405.01917*.

Mager, A. (2012). Algorithmic ideology: How capitalist society shapes search engines. *Information, Communication & Society*, 15(5):769–787.

Makhortykh, M., Urman, A., and Wijermars, M. (2022). A story of (non) compliance, bias, and conspiracies: How google and yandex represented smart voting during the 2021 parliamentary elections in russia. *Harvard Kennedy School Misinformation Review*, 3(2):online.

Markham, A. (2024). Algorithms as conversational partners: Looking at google auto-predict through the lens of symbolic interaction. *New Media & Society*, 26(9):5059–5080.

McDonald, N., Schoenebeck, S., and Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23.

Miller, B. and Record, I. (2017). Responsible epistemic technologies: A social-epistemological analysis of autocompleted web search. *New Media & Society*, 19(12):1945–1963.

Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., and Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10):1365–1386.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Olteanu, A., Diaz, F., and Kazai, G. (2020). When are search completion suggestions problematic? *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2):Article 171.

Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., and Ziegele, M. (2021a). Who is responsible for interventions against problematic comments? comparing user attitudes in germany and the united states. *Policy & Internet*.

Riedl, M. J., Whipple, K. N., and Wallace, R. (2021b). Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others. *Information, Communication & Society*, pages 1–18.

Robertson, R. E., Williams, E. M., Carley, K. M., and Thiel, D. (2025). Data voids and warning banners on google search. *arXiv preprint arXiv:2502.17542*.

Roy, S. and Ayalon, L. (2020). Age and gender stereotypes reflected in google's "autocomplete" function: The portrayal and possible spread of societal stereotypes. *The Gerontologist*, 60(6):1020–1028.

Schoenebeck, S., Haimson, O. L., and Nakamura, L. (2021). Drawing from justice theories to support targets of online harassment. *New Media & Society*.

Shim, Y. and Jhaver, S. (2026). Incorporating procedural fairness in flag submissions on social media platforms. *ACM Transactions on Social Computing*, 9(1):1–40.

Urman, A., Hannak, A., and Makhortykh, M. (2024). User attitudes to content moderation in web search. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–27.

Vaccaro, K., Xiao, Z., Hamilton, K., and Karahalios, K. (2021). Contestability for content moderation. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Wengraf, T. (2001). *Qualitative Research Interviewing*. SAGE Publications, Ltd, London.

West, S. M. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*.

Williams-Ceci, S., Jakesch, M., Bhat, A., Kadoma, K., Zalmanson, L., and Naaman, M. (2024). Bias in ai autocomplete suggestions leads to attitude shift on societal issues.

Zhang, A. Q., Montague, K., and Jhaver, S. (2023). Cleaning up the streets: Understanding motivations, mental models, and concerns of users flagging social media posts. *arXiv preprint arXiv:2309.06688*.