On the social and technical challenges of Web search autosuggestion moderation
by Timothy J. Hazen, Alexandra Olteanu, Gabriella Kazai, Fernando Diaz, and Michael Golebiewski

Abstract
Past research shows that users benefit from systems that support them in their writing and exploration tasks. The autosuggestion feature of Web search engines is an example of such a system: It helps users formulate their queries by offering a list of suggestions as they type. Such autosuggestions are typically generated by machine learning (ML) systems trained on a corpus of search logs and document representations. These automated methods can however become prone to issues that might result in the system making problematic suggestions that are biased, racist, sexist or in other ways inappropriate. While current search engines have become increasingly proficient at suppressing many types of problematic suggestions, there are still persistent issues that remain. In this paper, we reflect on past efforts and on why certain issues still linger by covering explored solutions along a prototypical pipeline for identifying, detecting, and mitigating problematic autosuggestions. To showcase their complexity, we discuss several dimensions of problematic suggestions, difficult issues along the pipeline, and why our discussion applies to an increasing number of applications (beyond Web search) that implement similar textual suggestion features. By outlining several persistent social and technical challenges in moderating Web search suggestions, we hope to provide a renewed call for action.

Contents
1. Introduction
2. Characterization
3. Discovery and detection
4. Difficult issues
5. Applications and future directions
6. Conclusions

1. Introduction
Web search query autosuggestion (often also referred to as autocompletion) is a feature enabled within the search bar of many Web search engines such as Google (Sullivan, 2018) and Bing (Marantz, 2013; Gulli, 2013). It provides instant suggestions of complete queries based on a partial query entered by a user [1]. Such query suggestions can not only help users to complete their queries with fewer keystrokes, but they can also aid them in avoiding spelling errors and, in general, can guide users in better expressing their search intents. Figure 1 shows an example set of suggestions by a search engine for the query prefix “search autoco.” Autosuggestion systems are typically designed
to predict the most likely intended queries given the user’s partially typed query, where the predictions are primarily based on frequent queries mined from the search engines past query logs (Cai and Rijke, 2016).

Figure 1: Top 4 example suggestions that were provided by Google’s search bar for the typed query prefix “search autoco”.

Despite its advantages, there are also pitfalls that can accompany this search feature. Since the suggestions are often derived from search logs, they can, as a result, be directly influenced by the search activities of the search engine’s users. While these activities usually involve benign informational requests, they can also include queries that others may view as loathsome, prurient, or immoral. If left unchecked, a wide variety of problematic queries may be mined from the logs including, but not limited to, queries that provide exposure to violent, gory, or sexually explicit content; promote behavior that is harmful, hateful, or illicit; or wrongfully defame people or organizations. The feature could also suggest more pernicious queries that expose undesirable biases and stereotypes, or spread misinformation; among many other issues.

The dependence on query logs also leaves the autosuggestion feature susceptible to manipulation through adversarial attacks. By coordinating the submission of a large number of specific queries, it is possible for an outside entity to directly manipulate the list of suggestions that are being shown for targeted query prefixes (Starling, 2013). For instance, two common scenarios where this happens are online promotion campaigns and intentional “trolling.” This issue is often enabled and further exacerbated by the so called data voids — topics searched in the past by little to no users and that reputable sources do not cover (Golebiewski and boyd, 2018) — due to a lack of competing queries for rare query prefixes (see §4.2).

The presence of highly offensive or derogatory suggestions within the query suggestions made by prominent search engines has not gone unnoticed. Numerous negative stories in the news media have highlighted problematic suggestions made by the various search engines (Elers, 2014; Lapowsky, 2018; Hoffman, 2018; Chandler, 2018). Real search autosuggestions were, for instance, also used in a high profile 2013 ad campaign by UN Women to highlight the “widespread prevalence of sexism and discrimination against women” (UN Women, 2013). The
campaign highlighted examples like ‘women shouldn’t have rights,’ ‘women cannot be trusted,’ ‘women should be slaves’ (where the part in square brackets represents a search suggestion that completes the partially typed query preceding it). Similar issues have also been highlighted by legal complaints concerned with suggestions allegedly defaming an individual or an organization (e.g., a suggestion implying the plaintiff is a ‘scam’ or a ‘fraud’) or promoting harmful illicit activities (e.g., a suggestion pointing to pirated versions of a plaintiff’s content) (Ghatnekar, 2013; Cheung, 2015; Karapapa and Borghi, 2015).

Sustained efforts to address these issues are critical to minimizing harms and maintaining public trust. Such efforts to reduce problematic Web search suggestions have resulted in noticeable improvements in recent years, though problems do remain. The actual mechanisms being used by particular search engines, however, have mostly remained hidden from the public view to avoid giving adversaries sufficient information to circumvent them. Thus, studies of issues surrounding autosuggestions tend to come from outside organizations that externally probe search engines using tools such as Google Trends or the Bing Autosuggest API (Diakopoulos, 2014), through qualitative audits (Tripodi, 2018; Diakopoulos, 2013b), or by programatically running queries, typically through dedicated browser plugins (R.E. Robertson, et al., 2019). However, due to the limited access to search logs, comprehensive studies by the broader research community are difficult to run. Even for those developing search engines, the sheer number and diversity of information needs, of search log queries, as well as of possible suggestions for each given prefix that the autosuggestion mechanisms may consider surfacing, makes the efficient discovery and detection of (both broader types and particular instances of) problematic query suggestions extremely difficult.

**Figure 2:** Organization and discussion overview.

**Organization & contributions.** In this paper, we provide an overview of a range of social and technical challenges when moderating Web search autosuggestions. Our goal is to highlight difficult and on-going issues in Web search autosuggestions that are still prevalent, and that require deeper investigation and potentially new methodologies in order to solve. We start by defining what it means for a query to be “problematic” and broadly characterizing several types of problematic query suggestions that can be surfaced (§2), which we then follow with a discussion about methodologies for discovering problematic queries within large volumes of queries considered by an autosuggestion system (§3.1). We then cover technical difficulties in designing automatic methods to accurately detect and suppress problematic queries at scale (§3.2). We follow this with a discussion of difficult editorial decisions that must be considered when designing a suggestion moderation system (§4). We conclude with a brief overview of similar applications the issues we cover here extend to, and highlight a few key research directions (§5).
2. Characterization

Identifying and mapping the universe of problematic query suggestions is difficult both due to the open domain, long-tailed nature of Web search, and variations in how people appraise the potential for harms. In fact, there is no standard definition of what and when suggestions should be considered problematic, with existing efforts largely focusing on some well agreed upon classes of problematic suggestions such as adult content, hate and illegal speech, or misinformation.

Many of these classes, however, are themselves ambiguous with no agreed upon definition. Take the case of hate speech which is codified by law in many countries, yet the definitions vary widely across jurisdictions (Sellars, 2016). Even for adult content there are debates about what should be included and what is harmful, and it is often challenging to make the call without proper historical or cultural context. One such example is the iconic Napalm Girl picture, which Facebook censored and then reinstated due to public criticism (Ibrahim, 2017).

**Working definition.** For our discussion here — to incorporate aspects of problematic suggestions mentioned by prior work, e.g., Diakopoulos (2013b); Cheung (2015); Yenala, et al., (2017); Elers (2014) — we follow Olteanu, et al., (2020) and broadly consider problematic any suggestion that may be unexpectedly offensive, discriminatory, or biased, or that may promote deceit, misinformation or content that is in some other way harmful (including adult, violent or suicidal content). Problematic suggestions may reinforce stereotypes, or may nudge users towards harmful or questionable patterns of behaviour. Sometimes third parties may also attempt to manipulate the suggestions to promote, for instance, a business or a Web site.

2.1. Dimensions of problematic suggestions

Understanding which suggestions should be construed as problematic and how to efficiently detect them also requires examining possible dimensions of problematic suggestions such as their 1) content (e.g., what type of content or topics are more likely to be perceived as problematic?) (Olteanu, et al., 2020; Miller and Record, 2017; Yenala, et al., 2017); 2) targets (e.g., who or what is more likely to be referenced in problematic queries?) (Olteanu, et al., 2020; Olteanu, et al., 2018; UN Women, 2013); 3) structure (e.g., are problematic queries likely to be written in a certain way?) (Santos, et al., 2017); and, 4) harms (e.g., what are the harms of surfacing problematic suggestions?) (Miller and Record, 2017).

While these dimensions do not necessary capture all aspects that make a suggestion problematic and other relevant dimensions could and should be considered, they help us illustrate the social, contextual, and topical complexity of providing Web search suggestions, as well as of their consequences to users and beyond. They also point to the need to design and implement discovery mechanisms and frameworks that can help capture a wider range of scenarios (§3).

2.2. Content

While existing work has mainly focused on the content of the queries, such as recognizing clearly racist, sexist or profane content (e.g., Yenala, et al., 2017), other types of problematic suggestions like subtle stereotypical beliefs have often been overlooked. We are interested in any topics and content categories which if expressed through a suggestion, then that suggestion is likely to be problematic. To illustrate the range of problematic suggestions, we highlight a few high-level categories that have been mentioned in prior literature, briefly discussing possible intersections and peculiarities. Given the long tail of problematic scenarios, however, in practice a different or more granular delimitation of the types of problematic categories might be needed (Olteanu, et al., 2020).

**Harmful speech.** A query suggestion may constitute harmful speech if the suggested query contains profane language (e.g., ‘<person> is a bastard’) [2]; if the query could be perceived as hateful, as it offends (e.g., ‘<person> is dumb’), shows a derogatory attitude (e.g., ‘girls are not that smart’) [6], intimidates (e.g., ‘refugees should not be allowed here’), or promotes violence towards individuals or groups (e.g., ‘kill all <-group>’); or if the query appears related to e.g., defamatory content promoting negative, unproven associations, or statements about individuals, groups, organizations (e.g., ‘<person> is a criminal’). In fact, these types of queries have been a focus of the prior work on the detection and suppression of problematic suggestions (P. Gupta and Santos, 2017; Diakopoulos, 2013a), as they often target individuals and groups (§2.3).
Illicit activities and speech. Query suggestions may also inadvertently nudge users towards illicit activities (e.g., ‘should I try [heroin]’) or may reproduce illicit speech (e.g., death threats like ‘<person> [needs to die]’) — such as appearing to promote terrorist/extremist content (e.g., ‘how to join [ISIS]’), or appearing to assert information that could be perceived as defamatory (e.g., ‘<person> [running a scam]’) (Diakopoulos, 2013a; Cheung, 2015). Because they often rely on past user queries, query suggestions may also inadvertently surface (and thus leak) private information (e.g., ‘<person> home address [<address>]’, ‘<person> health issue [<hospital name>]’).

Manipulation, misinformation, and controversy. Other query suggestions might be problematic if they could be perceived as controversial (e.g., ‘abortions should [be paid through tax money]’), if they appear to promote misinformation or information that is misleading or controversial (e.g., ‘climate change [is not proven]’), if they could nudge users towards conspiracy theories (e.g., ‘pizzagate [is real]’), or if they seem manipulated in order to promote certain viewpoints or content, typically content about ideological viewpoints, businesses or Web sites (e.g., ‘you should invest [in dogecoin]’).

Stereotypes & bias. Query suggestions might also be problematic if perceived by some users or other stakeholders as discriminatory, racist (e.g., ‘blackface is [OK]’), sexist (e.g., ‘girls are [bad at math],’ ‘women are bad [managers]’), or homophobic (e.g., ‘lgbtq people [shouldn’t have kids]’); as validating or endorsing certain political views (e.g., ‘democratic socialism is [bad for the country]’) (Borra and Weber, 2012) or certain ideological biases about certain groups (e.g., ‘democrats are [evil]’); or if they reflect systemic biases, stereotypes or prejudice often against a group (e.g., ‘women have [babies for welfare]’ or ‘immigrants [steal jobs]’). Detecting biases in suggestions may also require observing systemic patterns in what is being surfaced in relation to different groups (e.g., men versus women). Thus, while some of these types of problematic query suggestions intersect with some extreme types of harmful speech, many could also be subtle and challenging to detect.

Adult content. Perhaps one of the most tackled types of problematic queries are those containing pornography related terms, or those that could nudge users towards adult, obscene or racy content (e.g., ‘how to find [porn videos]’, ‘naked [women]’, ‘women that talk [about <racy phrase>]’) (Diakopoulos, 2013b).

Other problematic categories. While these categories already highlight the many ways query suggestions can be problematic, other noticeable types of problematic suggestions can also include self-harm and suicidal content (e.g., ‘how to [kill yourself]’, ‘ways to [harm yourself]’), promoting animal cruelty (e.g., ‘how to kill [a cat]’) (Diakopoulos, 2013b), reminders of some type of traumatic events (e.g., ‘what to do when [someone dies]’), or content on sensitive or emotionally charged topics for certain groups (e.g., ‘what is it like [to lose a child]’).

2.3. Targets

Anecdotally, certain types of problematic queries are more likely to mention certain types of subjects or targets. In fact, by looking at the prior literature (on both problematic suggestions or offensive speech), we observe a focus on content categories that often entail specific individuals or groups as the target, such as pornography (e.g., ‘women [naked]’), hateful speech (e.g., ‘arabs should [be deported]’), or stereotypes (e.g., ‘men do not [cry]’) (P. Gupta and Santos, 2017; Santos, et al., 2017; Olteanu, et al., 2018; Davidson, et al., 2017). However, the targets of problematic suggestions are often much more diverse (Olteanu, et al., 2020), including activities (e.g., ‘cutting yourself is [stupid]’), animals (e.g., ‘should i kill [my cat]’), organizations (e.g., ‘mainstream media is [destroying america]’, ‘nasa is [a joke]’), physical or mental illnesses (e.g., ‘bipolar disorder is [a fraud]’, ‘cancer is [not real]’), businesses (‘macy’s is [running a scam]’), or religions (e.g., ‘religion is [stupid]’); among many others.

2.4. Structural features

Discussed in more detail in the next section (§3), common templates can be observed in certain types of problematic queries, and many computational approaches to detecting such queries leverage a variety of linguistic cues. Indeed, it appears that structural features — such as how queries are being formulated (syntax, terminology, or speech acts), how long the queries are, and how specific they are — might correlate with some types of problematic suggestions (Davidson, et al., 2017; Santos, et al., 2017).

For instance, queries expanding with problematic suggestions may be structured and formulated as a sentence (declarative), question (interrogative), or only as a set of terms with no grammatical structure (e.g., ‘girls pics young’). Queries can also include different types of speech acts (e.g., assertive like ‘young people [are crazy]’ versus expressive like ‘upset that their [life is about to end]’), and can exhibit different levels of specificity (e.g., ‘drug dealers’ versus ‘drug dealers in new york city’) and can have varying lengths (e.g., a few terms versus full
On the social and technical challenges of Web search autosuggestion moderation

sentences). Such structural properties might also correlate with the demographic attributes of those writing them (Weber and Castillo, 2010; Aula, 2005), as well as with content (§2.2) and target categories (§2.3).

2.5. Harms

When determining whether a suggestion is problematic, the potential for various harmful effects — along with their severity, frequency or impact (Boyarskaya, et al., 2020) — should also be factored in (e.g., discomfort versus physical harm). Beyond the legal and public relation issues that might be caused by problematic suggestions (e.g., Karapapa and Borghi, 2015), search engines should thus mitigate a much wider range of potential harms. Miller and Record (2017) stress that search suggestions might induce “changes in users’ epistemic actions, particularly their inquiry and belief formation,” which can have harms like “generating false, biased, or skewed beliefs about individuals or members of disempowered groups.” Problematic suggestions could, thus, have a variety of consequences on both individual users and society.

At an individual level, suggestions may nudge users towards harmful or illicit patterns of behaviors (e.g., ‘should I try [heroin],’ ‘download latest movies [torrent]’), may offend (e.g., ‘women are [evil],’ ‘immigrants are [dirty]’), or may arouse memories of traumatic experience resulting in emotional or psychological harm (e.g., ‘my wish is [to die].’ ‘self-harm is [selfish]’). Suggestions referencing a specific individual can also be harmful if false or sensitive information is suggested about them, which at extremes can even constitute privacy breaches or can result in defamation.

Other types of harmful effects might also be caused by query suggestions that reinforce stereotypical beliefs (e.g., ‘girls are [bad at math]’); inadvertently promote inaccurate or misleading information (e.g., ‘bipolar disorder is [fake],’ ‘vaccines are [unavoidably unsafe]’) or particular norms, values or beliefs over others (e.g., ‘abortion should be [illegal]’); surface denials of or controversial stances about historical events (e.g., ‘hitler was [right]’); or appear to promote violence towards individuals, groups, animals or in general (e.g., ‘how to poison [my cat],’ ‘how to beat [my boss],’ ‘women should be [punished]’), among others.

3. Discovery and detection

The implementation of a system to suppress problematic queries typically involves two key stages. First, there is a discovery stage where example queries are reported by users, or mined from search logs and assessed as problematic or not (§3.1). This process typically generates a collection of annotated examples of problematic queries. Using this collection of queries, the second stage often includes the design and training of machine learning (ML) models for detecting problematic queries in order to suppress them from appearing as suggestions (§3.2).

3.1. Discovery of problematic scenarios

A key step in mitigating problematic query suggestion — though surprisingly under-explored by prior investigations and discussions about tackling various types of problematic suggestions — is their discovery. Although known methods for discovering problematic queries are varied, they are often ad hoc in nature, usually involving a combination of human intuition — typically of system designers about what is important and should be detected — with some sort of automated detection methods. This combination of intuition and automation can however leave them prone to important blind spots. Below we overview some of the approaches to detecting problematic query suggestion scenarios, and discuss challenges and some of their blind spots.

3.1.1. User reporting

Search engines often provide mechanisms for users to report problematic suggestions [2]. User reporting is invaluable to surfacing problematic suggestions that might have been missed by more automated detection mechanisms (discussed below), and typically results in the timely removal of the specific suggestions being reported. The resulting fixes, however, are reactive in nature and may be skewed towards the most salient cases, overlooking more subtle, lesser known, or cultural or context sensitive scenarios.

3.1.2. Red teaming
While user reporting is thus helpful, search engine providers prefer to discover problematic queries proactively and remove them before they are observed by any users. For this, search engine providers may employ “red teams” — i.e., teams of independent workers tasked with probing a search engine for various types of problematic suggestions. Red teaming seeks to expose weaknesses or blind spots in the search engine’s detection mechanism or even new classes of problematic suggestions that the system designers and developers were unaware of. This process can mimic white hat or ethical hacking practices (Caldwell, 2011; Palmer, 2001), where individuals can specialize in identifying such “vulnerabilities.” While red teaming could help proactively identify problematic query suggestion scenarios, it however remains challenging to scale.

3.1.3. Exploring common templates

After an initial discovery phase from user reports or red teaming exercises, known examples of problematic queries could be explored to identify common templates that frequently carry derogatory suggestions. For instance, template based approaches that leverage the syntactic structure of queries for collecting and sampling search data could help identify hateful and offensive speech with higher precision (Davidson, et al., 2017; Gitari, et al., 2015; Silva, et al., 2016); like the template ‘I <intensity><user intent><hate target>’ used by Davidson, et al. (2017) to identify targets of hate speech (e.g., ‘I really hate educated women’). Even simple query prefixes like ‘is <person> […]’ or ‘<person> is […]’ might often be completed with derogatory phrases. By collapsing queries into these common templates and then annotating the high frequency templates, additional derogatory phrases used to describe people could potentially be discovered more quickly. While template based approaches can scale, they will likely miss many queries that do not match the templates.

3.1.4. Query embedding

To help discover novel query forms expressing the same semantics as known problematic queries, deep-learned semantic embeddings can also be used. For instance, query embeddings can be learned from Web search click data to create a vector embedding space in which queries with similar click patterns are placed close together in the embedding space (Huang, et al., 2013; Shen, et al., 2014). By exploring queries located close to known problematic queries in an embedding space, new problematic query formats and derogatory phrases could be discovered. However, this is sensitive to the seed set of known problematic queries, and will likely miss cases that differ in nature.

3.1.5. Active learning

To take advantage of both machine learning models and human effort, an active learning approach could also be taken. Active learning is a human-in-the-loop method where machine learned detection models (such as those we discuss later in §3.2.3) can examine large volumes of unannotated queries and propose the queries most likely to improve the model if annotated (Settles, 2009). Typically these are borderline queries in the model’s score space that the model has high uncertainty on. By iteratively repeating the annotation of new queries proposed by the model and then retraining the model with the new data, both new problematic queries can be discovered and the ML detection models can be improved.

3.1.6. Linkage to problematic Web sites

The search results returned for a given query can also be used to assess a query’s propensity to be problematic. If the returned results are themselves in some way problematic (e.g., they contain adult content, or they are of low quality), the query itself may also be problematic (Parikh and Suresh, 2012). Independent of the actual content of a query, a search engine should not suggest any query that if selected would return as search results problematic Web sites. This could include, for instance, sites whose content is pornographic, incites violence, promotes illegal activity, or contains malware. Because of the potential for black hat adversaries to manipulate search results for particular queries (including those of a seemingly benign nature), discovering such queries may require the detection of problematic Web sites in the search results returned for the query (Lee, Hui, and Fong, 2002; Arentz and Olstad, 2004; G. Wang, et al., 2013).

3.2. Detection & suppression methods

Because of the open nature of language, the complexity of the problem, and the potential harm of surfacing problematic queries, search engines typically employ a mix of multiple manual and automated methods for detecting and suppressing problematic queries in order to improve recall and robustness (Santos, et al., 2017). Some of the
most common approaches are discussed below.

3.2.1. Block lists

It is common for systems to maintain manually curated block lists of highly offensive terms that will trigger the suppression of a suggestion that contain those terms. However, many terms are offensive only in certain contexts, e.g., scum is an offensive suggestion in ‘<person> is [scum]’ but non-offensive in ‘what is pond [scum]’. Therefore, detection techniques need to model the entire query to avoid erroneously suppressing legitimate query suggestions.

Block lists for whole queries can also be employed, and can be implemented efficiently for use in a run-time system. These lists allow for on-the-fly updating if a content moderator needs to immediately suppress an offensive query suggestion reported by a user. Block lists also ensure previously flagged queries remain permanently suppressed even if other modeling techniques unexpectedly leak the query after a model update. However, block lists cannot be generalized easily and (as with common templates) are not a scalable solution if used alone.

3.2.2. Query templates and grammars

Because search queries tend to be short and the likelihood of query uniqueness increases with query length, the collection of common query candidates used by an autosuggest feature will be dominated by short queries [10]. This makes it possible to sufficiently cover the head of the distribution of derogatory queries with simple templates or finite state grammars, though this approach is not practically feasible for covering the tail.

For example, to suppress derogatory suggestions about named individuals a simple approach is to use an entity extractor to identify people’s names within queries, curate a list of derogatory expressions, and then identify the most common query templates that use a combination of a person’s name and a derogatory expression such as ‘<person> is <derogatory_expression>’, ‘is <person> <derogatory_expression>’, or ‘why is <person> so <derogatory_expression>’.

To cover the wide range of template variations, a more efficient approach is to encode them into a rule-based grammar using a regular expression based finite-state tool (e.g., Foma [Hulden, 2009]). This has the advantage of quickly generalizing the head of the derogatory query distribution with high precision, but can still suffer from poor coverage for longer or less-common queries. It also typically require a human to update the grammar, which is not scalable for covering tail queries and improving recall.

3.2.3. Machine learning models

To achieve greater generalization and avoid the use of hand-written rules, ML approaches can also be used to detect problematic queries. Techniques that have been explored for this purpose include gradient boosting decision trees (Chuklin and Lavrentyeva, 2013), long short-term memory networks (Yenala, et al., 2017), and the deep structured semantic model (P. Gupta and Santos, 2017). Relative to grammar-based models, ML models have been found useful in improving the detection rate of problematic queries but at the expense of increased false positive rates (P. Gupta and Santos, 2017). While ML models avoid the effort required to hand-craft rules, they require human effort to annotate collections of queries (both problematic and non-problematic) in order to train models.

3.2.4. N-strike rule

One scenario that can plague a query suppression mechanism occurs when a system successfully removes problematic queries from a suggestion list only to have them replaced by other problematic suggestions missed by the detection model. This situation is even more likely when adversarial attacks specifically seek to find holes in the model. One way to combat this is to employ what is dubbed as the N-strike rule, i.e., the detection of N or more problematic queries at the top of a pre-filtered suggestion list for a query prefix will trigger the suppression of all suggestions for that prefix.

While it may be unclear for any single, arbitrary example what mechanisms were used to suppress suggestions, the complete suppression of suggestions for certain problematic prefixes is observable in both Google and Bing. For example, at the time of this writing, neither search engine provided any suggestions for prefixes such as ‘jews are,’ ‘muslims are,’ and ‘catholics are,’ which, without such interventions, did yield multiple highly inappropriate suggestions in the past (Gibbs, 2016).
4. Difficult issues

There are still lingering challenges to mitigating problematic query suggestions that require additional research, maybe even re-thinking current approaches to characterizing, discovering, detecting, and suppressing these suggestions; challenges we cover next.

4.1. Setting the boundaries of “problematic”

4.1.1. Operationalizing “problematic”

Even with a clear definition of what constitutes problematic query suggestions, setting the boundary between problematic and non-problematic cases can still be difficult. For example, certain system provided suggestions may be deemed offensive by some people but not by others.

A typical way to determining whether query suggestions are problematic is through some form of crowd labelling or by using block lists (for both data collection and data annotation). Both approaches however have limitations when used to operationalize ambiguous, latent concepts like stereotyping (Blodgett, et al., 2021). For instance, there are many factors that can affect human assessments including ambiguous definitions, poor understanding of the concepts, poor annotation guidelines, poor category design, or insufficient context (Olteanu, et al., 2019; Blodgett, et al., 2021). For concepts like hate speech, the characteristics of the users or of the crowd judges (e.g., their demographics or experiences) can also lead to important variations in their assessments of what constitutes hate speech, even when provided with the same definitions and instructions (Olteanu, et al., 2017).

4.1.2. “Problematic” in context

Another issue is that the context in which a suggestion is surfaced in can also impact how that suggestion is being perceived. In fact, there are cases where the suggestion might be problematic because of the context (or the lack thereof context). Consider the examples: ‘why do teens [commit suicide],’ ‘can you purchase [a gun online],’ or ‘what does the [bible say about abortion].’ Such suggestions might be problematic since what the user wrote (the query prefixes ‘why do teens’, ‘can you purchase’, or ‘what does the’) did not warrant the surfacing of sensitive or controversial topics. Getting the context right in these cases however is often hard.

Furthermore, how various types of problematic suggestions are expressed can vary across contexts and over time. In addition to active forms of obfuscation to trick the engine detection and suppression components (§4.2), the way in which, for instance, harmful speech is expressed is not always static and predictable, but can vary based on the identity of the author and that of the target, based on the time of the day, or based on some world events (Schmidt and Wiegand, 2017; Cheng, et al., 2017; Kumar, Cheng, and Leskovec, 2017), among other factors.

4.1.3. Subjectivity

Personal beliefs about topics such as religious beliefs, gender identity, abortion, and immigration, can also contribute to the subjective perceptions of what is problematic. For example, a query suggestion of the form ‘<person> is [gay]’ may be perceived differently by various people depending on, for instance, their own personal views, their prejudices and biases, and who the target of the query is. Someone with strong anti-LGBTQ beliefs might perceive this suggestion as problematic, especially if they are the person referenced in the query. The perception of such query suggestions can thus vary across the general population, e.g., the term gay is not offensive and it is commonly used to convey sexual identity (GLAAD, 2011), but has also unfortunately been used with an intolerant pejorative intent (Winterman, 2008). These differences in underlying intent can be nuanced and difficult to distinguish without proper understanding of the context. Consideration should be given to how assessments of what is problematic should be handled, and whether e.g., a ‘majority’ agreement (often the standard for crowdsourcing assessments) is an appropriate criteria for flagging suggestions as problematic. This is particularly critical if some of those doing the assessments lack the experience, the proper context, or the cultural sensitivities to make the correct judgement.

4.1.4. Truthfulness & mislabeling

A key driver for suppressing suggestions is preventing defamatory statements from appearing as suggestions. For
example, a query like ‘serial killer [<person>]’ could be viewed as defamatory if the person is not a serial killer. If the person is verifiably a serial killer, then the query is not defamatory and could be a common legitimate query that should not be suppressed. Avoiding suppression in such cases should require verification from a trusted knowledge base before showing the query.

A pernicious form of untruthfulness is intentional mislabeling. As described by Molek-Kozakowska (2010), intentional mislabeling can be a political ploy for “introducing and/or propagating terminology that is either inaccurate or derogatory (or both) to refer to a person, a group, or a policy in order to gain political advantage.” Such mislabeling is often used to imply someone is different from what they proclaim to be, or in order to suggest a difference that is perceived negatively by the intended audience of the messaging. A high profile instance is the mislabeling of Barack Obama as a Muslim by his political opponents (Holan, 2010).

MislABELing is particularly difficult to detect in autosuggestions because the terminology used is often not problematic in its own right, and would go undetected by typical filtering methods. Other mislabeling forms could involve e.g., political stance (e.g., ‘<person> is a [communist]’) and gender (e.g., ‘<person> is a [man]’).

4.1.5. Newsworthiness

In some situations, clearly offensive query suggestions may result directly from current newsworthy events, particularly when prominent individuals make derogatory statements about other people. In news media, editorial decisions are made to weigh the potential harm of reporting offensive content against the public’s need for full knowledge of an event. An example that news editorial teams struggled with was Donald Trump’s profanity laden reference to African countries (Jensen, 2018). Similarly, if an autosuggestion mechanism has the ability to discern when a query is referencing a current news story, it may also consider temporarily allowing a problematic query as an autosuggestion while the query’s subject matter is active in the news cycle. This approach would provide easier access to legitimate news stories while they are active, but suppress them when they fall out of the news cycle. However, doing so while ensuring that proper context is available for the query suggestion, is often difficult.

4.1.6. Historical inquiry

In general, negative queries about long deceased persons are unlikely to be submitted for the intent of defaming the person, and are thus more likely submitted for the purpose of historical inquiry. In cases where a historical figure is involved (e.g., ‘was woodrow wilson [a racist]’), the suppression of such query suggestions is likely unnecessary. Again, allowing such suggestions would typically require verification that the target is a deceased person of historical importance from a trusted knowledge source.

4.2. Challenges for detection methods

Even with a good understanding of what constitutes problematic scenarios, detecting abusive or other types of problematic language can be difficult. In some cases, they can be subtle, fluent, formal, and grammatically correct, while in others they can be ambiguous and colloquial (Nobata, et al., 2016). Other challenges to detecting problematic query suggestions involve adversarial queries, data voids, and others; which we discuss next.

4.2.1. Adversarial queries

As in any adversarial situation, people intending to corrupt autosuggestions for their own purposes will probe for novel ways to defeat any algorithmic suppression model. Adversaries may use a variety of tricks to mask the intent of a query from automated detection while still providing a clear intent to a typical user. A common circumvention strategy is to rewrite an offensive or problematic query to include misspellings, abbreviations, acronyms, homophones, leet speak, or other types of text manipulations.

Prior work particularly focused on hate and offensive speech had documented such strategies that circumvent abuse policies and automated detection tools. For instance, to avoid being detected by automated hate-speech detection tools, some users have developed a code (e.g., the operation Google Movement) in which references to targeted communities are substituted by “benign” terms in order to seem out of context (Magu, et al., 2017). Similarly, subtle changes applied to toxic phrases (e.g., “st.upid” instead of “stupid”, “idiiot” instead of “idiot”) have also proved effective in deceiving automated tools like Google’s Perspective; tools that are also susceptible to false positives (e.g., by not correctly interpreting negations) (Hosseini, et al., 2017).
By qualitatively examining search logs from Bing, we found similar examples including ‘<person> is an eediot,’ ‘<person> is a a hole,’ ‘<person> is a pos,’ or ‘<person> is a knee gar.’ Indeed, misspelled words — which are also less competitive in terms of the click-through rates and the traffic they draw — have become a target for manipulation, with adversarial parties employing increasingly sophisticated techniques to circumvent countermeasures that e.g., attempt to do automated corrections (Joslin, et al., 2019).

4.2.2. Data voids

The open ended nature of search allows users to look up anything and everything. Yet, some topics and their corresponding queries will be more popular than others — more users will search for ‘treating a cold’ than for ‘quantum computing’ or ‘asexuality.’ This leads to topics and associated queries for which there is little to no Web content. Instances where “the available relevant data is limited, non-existent, or deeply problematic” have been dubbed as data voids (Golebiewski and boyd, 2018). For instance, the anti-vaccine movement is believed to have leveraged existing information voids to promote their beliefs within the top results (DiResta, 2019). Similarly, there will also be queries (and query prefixes) that are less frequent, making suggestions much more prone to manipulation and to errors — e.g., such as misspelling of popular queries (Joslin, et al., 2019). One such example in Bing’s search logs is ‘<health insurance> of ill,’ where “ill” is a common misspelling of the Illinois state’s abbreviation (IL) (Olteanu, et al., 2020).

In fact, each query tends to be associated with its own unique demographic fingerprint (Shokouhi, 2013), with rare queries being frequently run by niche groups of users, and some of these groups can be adversarial in nature. Leveraging this at run-time however can be computationally intensive, and could raise both privacy and bias concerns (if e.g., queries from certain groups of users are more routinely suppressed).

4.2.3. Derogatory queries and trolling

Prominent public figures, such as polarizing politicians or partisan political commentators, might be particularly susceptible to the appearance of derogatory autosuggestions culled from user queries. While “trolling” of such individuals likely happens, it is often unclear whether any particular derogatory query suggestion is the result of a coordinated “trolling” attack or simply arises through organic searches submitted by users with negative views of an individual. For search engines that frequently update their candidates for query autosuggestions, negative news events can generate large query frequency spikes that could introduce new derogatory suggestions about such individuals.
Figure 3 shows query trends for two prominent and politically polarizing figures in the United States. Each plot shows the relative query volume for a neutral fact based query (in blue) versus a derogatory query (in red) over an 18-month time period. By examining the query volume for each individual, it is easy to observe spikes following events that puts them in the public eye. Furthermore, the volume of negative queries typically spikes and often exceeds the volume of neutral queries when news events paint them in a more negative light. In systems whose suggestions are ranked using query frequencies within a recent time window, highly frequent negative queries can replace more neutral queries in the suggestion list and can remain there while their observed frequency spikes remain within the analysis time window. Rapid changes in query volume about individuals can be cues for moderators to examine prefixes associated with them for potential trolling.

4.2.4. Cultural sensitivities

Sometimes the perceived offensiveness of queries may be limited to certain national, ethnic or cultural sub-populations. Without an awareness of the derogatory terminology or social sensitivities particular to a sub-population, it can be difficult for a system’s designers, developers or even data annotators to recognize some queries as problematic.

To provide an example, the phrase “black and tan” in the United States refers to a layered beer cocktail. However, in Ireland it refers to the military forces sent to Northern Ireland to suppress the Irish independence movement of 1920 and 1921. Because of the brutal tactics employed by the group, the phrase is primarily pejorative in Ireland (Bell, 2016). In designing models and policies for content moderation, it is prudent to avoid as much as possible a lack of diversity across different ethnic, gender, political, or cultural groups in the pool of people creating, maintaining, and governing such content moderation systems.

4.2.5. Stereotyping

Common ML algorithms learn features and frequent associations about words and phrases, but they are often not able to capture deeper insights about the social or cultural aspects of a query. It may thus be difficult for ML algorithms to make nuanced decisions about which queries might reinforce harmful stereotypes.

For example, consider the suggestion ‘girls should play [softball]’ versus the suggestion ‘girls should play [baseball].’ There is nothing inherently problematic about girls playing either softball or baseball, but the former suggestion can subtly reinforce the stereotype that girls should not play baseball (a sport historically played primarily by boys) while the latter can be an empowering statement to enable girls to disregard the stereotype. Research into discovering common gender stereotypes present in, for example, learned word embedding vectors examines this issue (Bolukbasi, et al., 2016), but these solutions are often sensitive to various implementation parameters, and solving this problem for the wide range of social and cultural stereotypes possible in autosuggestions remains an open, difficult problem.

4.2.6. False suppression

It is sometimes difficult to distinguish a derogatory query from a legitimate non-offensive query based only on the query’s words. Some non-offensive queries can appear offensive when viewed without deeper knowledge of the intent. Consider these example suggestions: ‘stupid girl [jennifer nettles],’ ‘judd apatow [sick in the head],’ ‘prince william [trash].’ Without additional context these suggestions may appear to be derogatory statements about the mentioned individuals. Yet, “Stupid Girl” is actually a song by Jennifer Nettles, “Sick in the Head” is a book by Judd Apatow, and the query ‘prince william [trash]’ likely references a trash disposal service in Prince William County, Virginia and not Prince William the individual. Such examples show that only examining a query surface form is sub-optimal and can result in legitimate queries being suppressed.

Additional mechanisms that take advantage of knowledge gleaned from prior search results or other query
understanding models can help mitigate such erroneous suppressions. For example, the non-offensive nature of the query examples mentioned above could perhaps be identified through the use of Web search entity linking and detection tools that match these queries against entities present in a knowledge base (e.g., titles of works of art, business names). Of course, references to works of art containing highly offensive or profane phrases in their titles can still be suppressed through the use of phrase block lists.

More broadly, dealing with ambiguity has been found challenging in other similar settings. For instance, sarcasm can be confused with abusive language (Nobata, et al., 2016), while ambiguity may originate from both the use of language and how various types of problematic suggestions might have been defined.

4.2.7. Promotional manipulation

Evidence exists that a cottage industry of black hat search engine optimization (SEO) businesses focused on manipulation of search engines’ autosuggestion feature has emerged. Their goal is to embed queries promoting businesses or products into search autosuggestion lists for key query prefixes (Vernon, 2015). These suggestions can be particularly problematic if the activities of the promoted business are illicit (e.g., promoting software installations containing malware). These manipulations can be difficult to detect based on the query alone as they are designed to resemble similar legitimate queries, though there are often common search templates within which these manipulations typically occur. One study estimated that roughly 0.5 percent of query suggestions covering 14 million common query prefixes exhibit evidence of SEO manipulation. These manipulations covered a range of business sectors including home services, education, legal and lending products, technology, and gambling (P. Wang, et al., 2018).

4.3. User perceptions of suppression

When to suppress, what to suppress, and how to suppress query suggestions remains a topic of debate, with users perceptions often varying widely.

4.3.1. Is suppressing suggestions censorship?

The suppression of query suggestions has been sometimes portrayed as a form a censorship in the press and in online blogs (Anderson, 2016; Wyciślik-Wilson, 2016). This argument assumes that user queries themselves represent a form of content and search engines that return autosuggestions are providing a platform for exposing common user viewpoints via the queries the users submit. The removal of derogatory suggestions about individuals has been particularly derided by some political figures for its potential to direct users away from negative information about their opponents (Akhtar, 2016). Search engines defend their practices by indicating that the suppression of some query suggestions does not prevent users from submitting the search queries of their choosing, nor does it alter the content that is returned for any of their search query (Yehoshua, 2016).

4.3.2. Avoiding bias

There is evidence that the general public feels that major technology companies are politically biased and intentionally suppress certain viewpoints (Tiku, 2018). Some have also suggested that suppressions within search autosuggestions are purposely favoring one political party over another (Project Veritas, 2019). Search engines defend their query filtering mechanisms by claiming they apply the same filtering policies regardless of the political affiliations of the individuals or groups mentioned in the queries (Akhtar, 2016).

That said, major search engines serve a wide range of users and should thus strive to avoid even the perception of bias in their content moderation. Because humans may innately possess unconscious bias against certain points of view or certain people, it is important to have mechanisms in place that minimize any biases that might originate from the process of annotating data and from designing or training ML models.

For example, a person’s political views may influence their perception of whether a query suggestion is problematic. Consider a suggested query of the form ‘<person> [is corrupt].’ this query template would generally be considered as a derogatory and potentially defamatory statement about the person. However, if we show this query form using a particular person’s name in place of the <person> marker, an annotator’s perception of the offensiveness of the query may be reduced or heightened based on their personal opinion of the mentioned person. Designers of annotation tasks should remove or account for the potential for such biases wherever possible.
Additionally, biases from a search engine’s user population can also seep into the moderation process (Olteanu, et al., 2019). This can, for instance, occur through the feedback mechanisms that allow users to report content or suggestions as offensive. If a search engine’s user base is e.g., skewed in favor of a particular political viewpoint, then suggestions deemed derogatory about politicians they agree with may be reported, reviewed by a moderator, and subsequently suppressed more frequently than suggestions about politicians with an opposing viewpoint. It is thus important for the moderation process not only to suppress the unique queries reported by users in an online block list (which is common for immediate suppression of offensive content), but also to update the detection model so the offensive query form will be suppressed for other individuals as well.

4.3.3. Introduction of positive bias

Even if care is taken to avoid bias during the moderation process, the presence of query suppression can become obvious to users and cause them to perceive a suggestion list as biased or as intentionally manipulated.

For example, when applying an approach where derogatory, defamatory or offensive queries are being suppressed, the suggestion list can become susceptible to a reverse form of positive bias. For a highly polarizing public figure, the list of queries submitted about that person will contain a mix of both highly positive and highly negative queries, in addition to neutral fact-oriented queries. In adversarial settings, their autosuggest list without any suppression would likely be dominated by highly negative queries like those in the “Before” column in Table 1. However, after removing the derogatory queries, their autosuggest list might end up looking abnormally positive in nature, such as in the “After” column in Table 1. While the positive queries that remain after suppression are not offensive in nature, they may appear abnormally biased in the positive direction given the polarizing nature of an individual. In such cases, a neutral approach where only queries that express neither a positive nor a negative sentiment are shown could help avoid the appearance of bias.

<table>
<thead>
<tr>
<th>Before Suppression</th>
<th>After Suppression</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;person&gt; is a liar</td>
<td>&lt;person&gt; is back</td>
</tr>
<tr>
<td>&lt;person&gt; is a joke</td>
<td>&lt;person&gt; is right</td>
</tr>
<tr>
<td>&lt;person&gt; is evil</td>
<td>&lt;person&gt; is awesome</td>
</tr>
<tr>
<td>&lt;person&gt; is a moron</td>
<td>&lt;person&gt; is king</td>
</tr>
<tr>
<td>&lt;person&gt; is wrong</td>
<td>&lt;person&gt; is great</td>
</tr>
<tr>
<td>&lt;person&gt; is dumb</td>
<td>&lt;person&gt; is from where</td>
</tr>
</tbody>
</table>

Table 1: Example of potential effects of suppression of derogatory queries for the top of an autosuggestion list for the prefix form “<person> is” both before and after the query suppression model is applied.

5. Applications and future directions

Throughout this paper, we have identified a variety of lingering issues with Web search query autosuggestions,
which point to several research areas that we believe require renewed attention. Among these, critical areas include tackling biases, computational harms and subjectivity, understanding the impact of moderation related interventions (or their lack off), and distilling the role of contextual cues (§5.1). Additionally, the issues we have described are more and more pressing as they apply to a growing number of similar applications, where textual suggestions are being offered to users (§5.2).

5.1. Key research topics

The issues surrounding bias and subjectivity are particularly difficult. There are instances where suggestions would not be deemed problematic by a majority of users, but could (justly) be viewed as highly problematic by a (small) minority (Olteanu, et al., 2020). This highlights the difficulties search engines face when defining and executing a policy about what is considered problematic, especially when the judgements on whether given queries are problematic are crowdsourced (as it is often the case). Where should the line be drawn on problematic queries, and how do decisions about this affect the users and their experience?

To avoid the appearance of bias, perhaps search engines should, as much as possible, avoid promoting specific opinions within their suggestions in favor of providing more neutral or objective suggestions. Yet, even when all candidate suggestions are neutral, bias could still creep in and result in certain types of associations or references being surfaced in certain contexts but not others (e.g., age might be more often associated with women than with men). How should candidate suggestions be mixed and varied across users, time, and contexts to avoid either obfuscating or promoting certain associations?

It is unclear the extent to which moderation of suggestions affects search engine utility or user satisfaction. Would users find suggestions more helpful if they were unbiased and focused on neutral information seeking? Or, do users prefer to see suggested queries submitted by other users even when they are biased, offensive or problematic in other ways? How should the impact of moderation-related interventions (or their lack thereof) be appraised? When suggestions are both useful (matching the user’s search intent) and problematic, should a search engine still surface them? [11] These are all open questions that require additional study to understand both the actual and perceived effects of suggestion moderation on Web search.

With misinformation rampant on the Internet, a major issue is whether suggestions contribute to its spread. Many suggestions may not appear problematic on their own, but can be problematic if they help promote misinformation, such as directing users toward unfounded conspiracy theories even when they are not looking for them (Roberts, 2016), or might do harm in other manners. To combat this might require deeper investigations into the Web sites that suggested queries point to and their authority (Hiemstra, 2020; Metaxa-Kakavouli and Torres-Echeverry, 2017).

Some contextual factors make search engines more prone to problematic suggestions. They can also affect how users perceived certain suggestions. Yet, it is often hard to both know which contextual factors to consider and how to distill their impact. Frequent queries are more likely to be suggested, but we may not know why they are frequent: do they reflect some prevalent needs or do they reflect attempts to manipulate the search engine? Furthermore, if autosuggestion feature relies only the prior queries by the current user, the user would perhaps be less likely to find the suggestions problematic. However, in many systems a general-purpose base model is still used as a backstop for contexts where a personalized model lacks data (Shao, et al., 2020), and thus problematic suggestions resulting from the base model may still occur even in scenarios with user personalization.

5.2. Extensions to other applications

Search query suggestions are part of a broader class of applications that provide text prediction or text rewriting suggestions to help users write faster, better, or in a more inclusive manner (Arnold, et al., 2018; Kannan, et al., 2016; Cai and Rijke, 2016; R.E. Robertson, et al., 2021). These applications aim to reduce the writing burden and help users efficiently complete tasks, and increasingly make use of large scale language models. However, they suffer from similar challenges to issue discovery and detection, as they operate in a similar fashion as the search autosuggestion engines do: using past writing samples (or usage logs) and limited context, they need to contribute suggestions for an open domain, long-tailed set of requests and needs. Such applications span Web search, e-mail and chat response and composition, as well as more general purpose assistive writing.

Other search applications. Besides providing suggestions as users type their queries, Web search engines also offer similar query suggestions via features like “Related searches,” “Top stories,” or “People also searched for.” While these features differ from search autosuggestions, as full queries or even topics are being suggested, their
appropriateness similarly depends on factors like cultural or historical context, or newsworthiness. As a result, these applications are similarly sensitive to societal biases, privacy breaches, data voids, and manipulation. Issues may also occur in enterprise search if, for instance, confidential or siloed information is being surfaced for users that should not have access to it.

**E-mail and chat responses.** The use of predictive text is also increasingly common in e-mail and other conversational environments. Relevant applications include the now popular Smart Reply (Kannan, *et al*., 2016) and Smart Compose (Chen, *et al*., 2019) features, which provide textual suggestions to users in the form of brief, standalone messages (for quickly sending a short reply) or as partial sentence completions (for composing longer response messages). While the systems generating these suggestions often leverage richer and cleaner data, they were also found to surface problematic suggestions (R.E. Robertson, *et al*., 2021) that might e.g., misgendered users (Vincent, 2018), or surfaced offensive associations (Larson, 2017). Due to their more personal use nature, these applications have also raised issues concerning the possibility of impersonating users (R. Gupta, *et al*., 2018).

**General purpose assistive writing.** More broadly, predictive text suggestions are used to assist users in any writing environment (Arnold, *et al*., 2016; Hagiwara, *et al*., 2019), including re-writing suggestions that make writing more concise and inclusive (BBC, 2019). Yet, due to the linguistic and social complexities of natural language that also affect the search query suggestions, such assistive writing systems can and have similarly failed to account for e.g., socio-cultural sensitivities (Scott, 2019).

**Applications of large neural language models.** The recent advances in large generative neural language models, such as GPT-3 (Brown, *et al*., 2020), and their use in a wide variety of text prediction tasks has also drawn particular attention of late. Because of the vast amount of data from a wide variety of sources used to train these models, they are also likely to reproduce biases, stereotypes and other social or cultural viewpoints that might be problematic. Users of these models are typically unable to curate the original data used to train the models and it is difficult to mitigate problematic issues contained in the training data once they are embedded into the model. The potential risks of using pre-trained large language models within applications that provide suggestions during text composition have been identified within in recent papers (McGuffie and Newhouse, 2020; Bender, *et al*., 2021).

6. **Conclusions**

We have highlighted some of the social and technical challenges that make the moderation of Web search autosuggestions difficult. While great progress has been made in suppressing the most obvious and harmful problematic suggestions, problems still remain as the query space has a long tail. We also caution that the implementation of similar features across other applications may raise new issues.

Many of the issues we covered throughout the paper are complex and difficult to mitigate. Some may even argue that these issues are so intractable that the safest approach is to disable the autosuggest feature entirely (despite its known benefits to users), or to at least allow users to opt-in to the feature with a warning about its potential problems. Alternatively, the feature could be used only to surface prior frequent queries made by the current user, which might preserve some benefit while eliminating problematic suggestions learned from other users. Others may also argue that the decision to moderate and the moderation processes are themselves problematic as they could reflect the biases of those tasked with moderating these systems (an observation sometimes used to argue for less rather than better moderation). For this paper, our goal was to review on-going existing issues posed by the search autosuggestion feature, with the hope that highlighting them will inspire new research and development efforts into the challenging aspects of the problems, both technical and social.

Greater transparency about the issues can also help alleviate some of the concerns about the practices of autosuggestion moderation. For example, Google has published a policy which provides a high level description of the types of queries that it suppresses [12]. Sustained efforts by search engine companies to improve the processes and deployed technologies will also help minimize harms and increase public trust in this generally helpful search feature.
About the authors

Timothy J. Hazen is a Senior Staff Machine Learning Researcher at Twitter. Direct comments to: thazen [at] twitter [dot] com

Alexandra Olteanu is a Principal Researcher at Microsoft Research Montreal. Direct comments to: aloltea [at] microsoft [dot] com

Gabriella Kazai is a Principal Applied Scientist at Microsoft Bing.

Fernando Diaz is a Staff Research Scientist at Google.

Michael Golebiewski is a Principal Program Manager at Microsoft Bing.

Acknowledgments

The authors have benefited from collaborations and interactions with a range of colleagues that deserve recognition. Thus, the authors wish to provide thanks for discussions and feedback to Eugene Remizov, Joshua Mule, Jose Santos, Balakrishnan Santhanam, Abhiyan Agrawal, Swati Valecha, Harish Yenala, Shehzaad Dhulaiwala, Zhi Liu, Keith Battochich, Toby Walker, Molly Shove, Marcus Collins, Mohamed Musbah, and Luke Stark.

Notes

1. We use the term autosuggestion instead of autocompletion as suggestions are not restricted to completions of a partial query, but can also allow alternative wording. These terms have also been used interchangeably throughout the literature, referring to scenarios where suggestions are based on the current user history, a subset of the queries issued in the past by all users, or a mix of both.

2. We are using this notation throughout the paper, i.e., ‘query prefix [suggestion]’.


5. To avoid disparaging any particular individual or group through our examples, we use place-holders like <person> or <group> as stand-in for any named person or group when the clarity of the query intent can be preserved.

6. Throughout the paper we showcase many illustrative examples for both specific types of problematic suggestions and for specific issues. While the examples have been edited for clarity and anonymity, they originate from either 1) search logs (not from existing autosuggestions); 2) an inventory of examples of autosuggestions encountered by the authors over time on various search engines; 3) prior published work; or 4) news media articles.

7. For instance, instructions for reporting offensive autosuggestions on Google can be found here: https://support.google.com/websearch/answer/106230. For Bing, feedback on offensive suggestions can be submitted by clicking the “Feedback” link on the bottom ribbon of the Bing page containing the problematic suggestion.


10. For instance, prior work found search queries to have an average length of about four to five words (R.E. Robertson, et al., 2019).

11. E.g., if a user believes that garlic cures cancer, should suggestions suggest that this is in fact true: ‘garlic [cures cancer]?’

References


On the social and technical challenges of Web search autosuggestion moderation


S. Ghatnekar, 2013. “Injury by algorithm: A look into Google’s liability For defamatory autocompleted search
suggestions,” Loyola Entertainment Law Review, volume 33, number 2, at

S. Gibbs, 2016. “Google alters search autocomplete to remove ‘are Jews evil’ suggestion,” Guardian (5 December),
at https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-
suggestion, accessed 14 July 2020.


Sengor Altngovde, D. Song, D. Albakour, S. Watt, and J. Tait (editors). Advances in information retrieval. Lecture
doi: https://doi.org/10.1007/978-3-319-56608-5_47, accessed 14 July 2020.


Processing and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP):
System Demonstrations.


C. Hoffman, 2018. “Bing is suggesting the worst things you can imagine” (10 October), at

A. Holan, 2010. “Why do so many people think Obama is a Muslim?” olitiFact (26 August), at
2020.

2020.

Web search using clickthrough data,” CIKM ’13: Proceedings of the 22nd ACM International Conference on
Information & Knowledge Management, pp. 2,333–2,338.


Y. Ibrahim, 2017. “Facebook and the Napalm Girl: Reframing the iconic as pornographic,” Social Media + Society
(20 November).


On the social and technical challenges of Web search autosuggestion moderation


M. Wyciślik-Wilson, 2016. “Google explains that search autocomplete censors suggestions,” at
On the social and technical challenges of Web search autosuggestion moderation


______________________________

Editorial history

Received 17 July 2020; revised 17 August 2021; revised 18 October 2021; accepted 25 January 2022.

This paper is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

On the social and technical challenges of Web search autosuggestion moderation
by Timothy J. Hazen, Alexandra Olteanu, Gabriella Kazai, Fernando Diaz, and Michael Golebiewski. First Monday, volume 27, number 2 (February 2022). doi: https://dx.doi.org/10.5210/fm.v27i2.10887