When Are Search Completion Suggestions Problematic?

ALEXANDRA OLTEANU, FERNANDO DIAZ, and GABRIELLA KAZAI, Microsoft

Problematic web search query completion suggestions—perceived as biased, offensive, or in some other way harmful—can reinforce existing stereotypes and misbeliefs, and even nudge users towards undesirable patterns of behavior. Locating such suggestions is difficult, not only due to the long-tailed nature of web search, but also due to differences in how people assess potential harms. Grounding our study in web search query logs, we explore *when system-provided suggestions might be perceived as problematic* through a series of crowd-experiments where we systematically manipulate: the search query fragments provided by users, possible user search intents, and the list of query completion suggestions. To examine *why* query suggestions might be perceived as problematic, we contrast them to an inventory of known types of problematic suggestions. We report our observations around differences in the prevalence of a) suggestions that are problematic on their own versus b) suggestions that are problematic for the query fragment provided by a user, for both common informational needs and in the presence of web *search voids*—topics searched by few to no users. Our experiments surface a rich array of scenarios where suggestions are considered problematic, including due to the context in which they were surfaced. Compounded by the elusive nature of many such scenarios, the prevalence of suggestions perceived as problematic only for certain user inputs, raises concerns about blind spots due to data annotation practices that may lead to some types of problematic suggestions being overlooked.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Information systems → Information retrieval**;

Keywords: web search suggestions, query logs, predictive text, problematic suggestions

ACM Reference Format:
Alexandra Olteanu, Fernando Diaz, and Gabriella Kazai. 2020. When Are Search Completion Suggestions Problematic?. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, CSCW2, Article 171 (October 2020). ACM, New York, NY. 25 pages. https://doi.org/10.1145/3415242

1 INTRODUCTION

There are many scenarios across computational media where assistive writing technologies are employed to aid users by providing suggestions that anticipate what they will write or do. Such assistive scenarios include web search query completion suggestions [8, 72] and predictive typing for email and chat response [5, 41, 70]. Powered by increasingly sophisticated natural language generation (NLG) models, the suggestion generation systems typically provide users with a small selection of suggestions. Such suggestions not only save typing effort and speed up repetitive tasks, but can also improve the quality of users' writing and that of their task's outcome—with web query suggestions also leading to better search results [8]. Though beneficial to users, these features have been scrutinized as reports of offensive and biased suggestions have emerged. One prominent example, a campaign run by United Nations' Entity for Gender Equality and the Empowerment of Women (UN Women) featured real offensive and demeaning suggestions for search queries about

Authors' address: Alexandra Olteanu, alexandra.olteanu@microsoft.com; Fernando Diaz, fdiaz@microsoft.com; Gabriella Kazai, gkazai@microsoft.com, Microsoft.

2573-0142/2020/10-ART171 $15.00
https://doi.org/10.1145/3415242

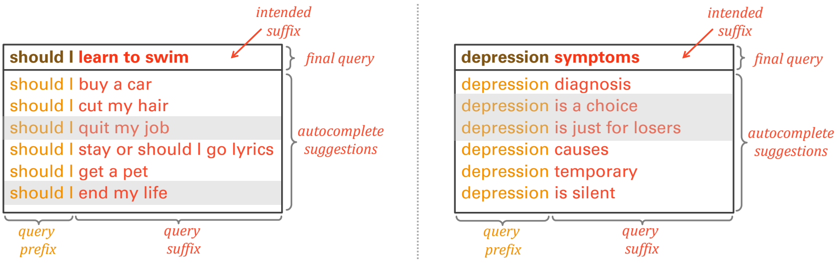


Fig. 1. Examples of search query completion suggestions for different query prefixes.

women [85]. Such problematic suggestions can have many pernicious effects, including reinforcing existing stereotypes and nudging users towards harmful patterns of behavior [36, 71, 85].

Problematic suggestions. Indeed, suggestions surfacing overtly racist or sexist views, or containing profanity or violence, have been extensively discussed in both media and research [12, 15, 16, 30, 31, 36, 43, 87]. However, suggestions can also be problematic in subtle ways, often dependent not only on the user input (the *query prefix*) and the provided query completion suggestions (*query suffixes*), but also on other contextual factors like the user's search intent (*intended suffix*) and socio-cultural peculiarities. A certain phrase may be bothersome for some group of users, but not for others. A suggestion such as "*... is evil*" or "*... is huge*" may be acceptable if the user input was "*heinous crime...*" or "*the universe...*," but likely problematic if it was a person's name.

Such context dependencies coupled with the long-tailed and open nature of search and writing tasks, make setting clear boundaries around the space of problematic suggestions impractical. Even if such boundaries could be drawn, the evolving nature of many contextual cues and the sheer diversity of factors that lead to suggestions being perceived as problematic still make it difficult to anticipate and enumerate all problematic scenarios.

Contributions. For system designers, however, a more systematic understanding of *when and why system-provided suggestions might be problematic* could help preempt future issues that they do not yet know to look for. Towards this exploratory goal, we developed a framework to help surface, identify, and characterize problematic suggestions (§3.1), in terms of both 1) what factors might lead to their exposure to users and 2) why they could be perceived as problematic by users. Given the context dependence and elusive nature of many problematic query completion scenarios, in this work we cast a wide net for what constitutes problematic suggestions (§3.2).

To conduct our study in the realm of web search, we based our investigation on query log samples from a large commercial search engine. Then, through a mixed-methods approach (§3.1) blending heuristics for query data sampling (§4.1) and synthetic query suggestion generation (§4.2) with crowd experiments (§5) and exploratory data analysis (§6), we examined *when problematic scenarios arise*. To further understand *why* suggestions are deemed problematic, we contrasted observed scenarios with a multi-dimensional inventory of known categories of problematic suggestions (§3.2).

Results overview: Overall, our results suggest a rich array of scenarios where suggestions are perceived as problematic, blind spots due to data annotation practices, and factors that make search engines more prone to surfacing problematic suggestions. Among others, key findings include:

- *Accounting for context:* Across our experiments, 15% to 47% of problematic suggestions were flagged as problematic due to the query prefix they were surfaced for (§6.3). This raises concerns about possible blind spots, as many of these scenarios are likely overlooked when queries are assessed without making a distinction between the prefix and the suggested completions—*the* common practice, e.g., see [20, 30].

- *Search voids:* In our experimental samples, we also found rare query prefixes to be up to 3 times more likely to be linked to problematic suffixes, corroborating anecdotal evidence that *search voids*—topics searched by only few users [26]—make suggestion engines more prone to problematic suggestions. Inspecting rare search scenarios could thus help forestall problematic suggestions before being surfaced to users.
- *Characterizing problematic scenarios:* More generally, we find query prefixes to have varying propensities to surface problematic scenarios, and interplays between *what* is referenced in suggested queries and *why* the suggestions are deemed problematic (§6.3–§6.4).

2 BACKGROUND AND PRIOR WORK

We are interested in scenarios where system-provided suggestions for web search query completions could be perceived as problematic by users, and to investigate factors that may affect the propensity of observing such scenarios. To design experiments that reflect actual characteristics of suggestion generation engines and of how users interact with them, we review prior work on query suggestion generation and query log analysis. Given that query suggestions are typically generated from behavioral patterns observed in past search logs [35, 47, 77], we also overview work that provides cues about how past logs can make suggestion engines more prone to surfacing problematic suggestions, and that can help us catalogue such cases.

2.1 Search Query Completion Suggestions

Whether generated using a machine learning (ML) system or some other algorithm, the selection of query completion suggestions for a given query prefix often depends on the frequency of query suffixes in past search logs [4, 8]. We also rely on query frequency in past logs to sample prefixes and generate synthetic query suggestions for our experiments (§4), to examine the impact of different distributions of prefix-suffix pairs on surfacing suggestions that are being perceived as problematic.

Prefix attributes. Studies of user interaction with search query completion suggestions found users most likely to engage with the feature at word boundaries [44, 51]. As a result, in our experiments we consider query prefixes and suffixes at word granularity (§4.2). Mitra et al. [51] also found users more likely to engage with the suggestions after typing out about half of their query, while Kharitonov et al. [38] that the longer the prefix the more likely users were to use the suggestions. Eickhoff et al. [21] estimate queries in sessions where users search for procedural knowledge (i.e., how to do something) or for declarative knowledge (i.e., find out about something) have a mean length of about 5 terms. These studies informed our query data sampling and processing (§4.2).

Search intent & suffix attributes. Query suggestion engines try to predict the most likely user intent for a prefix based on the prevalence of prefix-suffix pairs in query logs [47, 77]: e.g., a common suffix for the "*gmail ...*" prefix is "*... login*." Given the long tail of unique queries, though, most prefixes occur infrequently. Mitra and Craswell [50] describe a strategy to generate synthetic suggestions for rare query prefixes based on popular suffixes observed in the logs, which we take inspiration from to generate synthetic suggestions in our experiments. Users engagement with suggestions also depends on their search intent. Mitra et al. [51] observed a higher probability of engagement for how-to and navigational queries, hypothesizing this to be due to search engines being more likely to return relevant suggestions for popular queries. For a given prefix, we also experiment with different possible search intents to see if they affect how suggestions are perceived.

2.2 Problematic Query Completion Suggestions

Search queries may contain biased [3, 6, 37, 49, 54], racist or violent [22, 64, 76, 87], adult [15], or defamatory content [12, 16, 68]. Surfacing parts of such queries as completion suggestions to users can have negative implications if they offend, promote stereotypes, or even suggest the companies offering the service condone such views [12, 16, 30, 68]. While modern search engines are increasingly proficient at identifying unambiguously objectionable suggestions, *more context dependent scenarios remain a challenge.* Wang et al. [82] examines how spammers target query suggestion engines to illicitly promote certain associations, including by exploiting the engines' reliance on the observed popularity of queries in past logs. Their successful use of NLP techniques also suggests that manipulated suggestions might exhibit distinctive linguistic patterns. Shokouhi [72] also shows differences in the popularity of query suffixes across demographics, helpful for personalizing query suggestions. These suggest variations in how users assess suggestions, which we account for when aggregating crowd assessments (§5).

Auditing query suggestions. Prior work studied query suggestions by submitting partial queries to a search engine and collecting the resulting suggestions [10, 53, 67]. Such efforts also focus on a few types of suggestions that are, e.g., hateful or political, and their analyses are bounded to suggestions observed when partial queries are submitted. We do not only aim to assess suggestions that search engines are currently surfacing, being instead interested in broader "what if" scenarios—examining *what factors lead to suggestions being perceived as problematic* and *when are search engines prone to surfacing such suggestions.* Our experiments thus use logged queries, typically used to generate suggestions. We do however draw inspiration from these studies when assembling our inventory of known types of problematic suggestions (§3.2). Other efforts leverage advancements in ML to identify inappropriate queries [30, 87], but remain limited to a narrow range of issues (§6.3).

Ethical, social, and legal considerations. Due to the sheer volume of data they have to cull through, search engines like Google or Bing often employ content filtering and moderation strategies that depend on automation to scale up [9, 25]. These strategies are not error free, being often limited by the query examples they were trained on. Failing to identify problematic scenarios, however, may harm not only those issuing a search query, but also the *target* of the query e.g., when an individual's dignity may be compromised by suggestions surfacing false information about them [49]. Miller and Record [49] argue that certain suggestions should not be surfaced, like those perpetuating damaging stereotypes about socially disadvantaged groups or those associating negative characteristics with specific individuals. Legal controversies and liability may arise e.g., when query suggestions associate offensive or defamatory meanings, or could constitute infringement of intellectual property rights [36, 64, 75]. Such insights on the social and legal implications of system-provided suggestions also inform our inventory of known categories of problematic suggestions (§3.2).

3 EXPERIMENTAL DESIGN

We examine how user and query-related factors—like the user-provided query prefix and their search intent—lead to variations in the prevalence and type of problematic suggestions. To tease such factors apart, we explore three dimensions of problematic suggestions: 1) *context*—whether a suggestion is perceived as problematic in the context of a given query prefix typed by the user, as opposed to being problematic regardless of the prefix; 2) *query content or search topic*—what topics or types of content get mentioned in suggested queries perceived as problematic; and 3) *query target or subject*—who or what is referenced in such queries. To capture these dimensions, our experiments and exploratory data analysis are guided by the following set of assumptions:

- ***User input:*** *The query prefix typed by users in a search engine is a key factor in surfacing suggestions that are likely to be perceived as problematic.* In fact, we further conjecture that there are certain classes of prefixes for which the suggestions are more likely to be problematic. For example, a prefix like *girls should ...* or *abortion is ...* might be more prone to problematic suggestions.

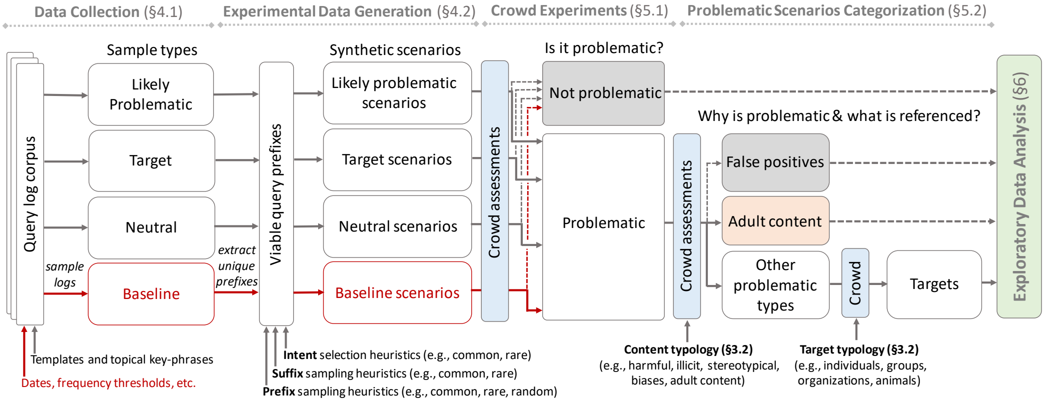

Data Collection (§4.1)
Experimental Data Generation (§4.2)
Crowd Experiments (§5.1)
Problematic Scenarios Categorization (§5.2)
Query log corpus
Sample types
Likely Problematic
Target
Neutral
Baseline
sample logs
extract unique prefixes
Viable query prefixes
Synthetic scenarios
Likely problematic scenarios
Target scenarios
Neutral scenarios
Baseline scenarios
Crowd assessments
Is it problematic?
Not problematic
Problematic
Why is problematic & what is referenced?
Crowd assessments
False positives
Adult content
Other problematic types
Crowd
Targets
Exploratory Data Analysis (§6)
Templates and topical key-phrases
Dates, frequency thresholds, etc.
Intent selection heuristics (e.g., common, rare)
Suffix sampling heuristics (e.g., common, rare)
Prefix sampling heuristics (e.g., common, rare, random)
Content typology (§3.2)
(e.g., harmful, illicit, stereotypical, biases, adult content)
Target typology (§3.2)
(e.g., individuals, groups, organizations, animals)

Fig. 2. Overview of our experimental framework.

- ***Search intent:*** *Knowing what a user wants to search for plays a minor role in the perception of suggestions as problematic by third-party judges.* In other words, even if users have some "dubious" search intents—e.g., suggestions could be both problematic and useful at the same time—the suggestions might still be perceived as problematic, and thus should remain neutral.
- ***Search voids:*** *Rare search query completion scenarios are more likely to expose users to problematic suggestions.* Low frequency queries are more susceptible to manipulation by adversaries, and are more frequently associated with lower quality data [26].
- ***Content properties:*** *Who or what is referenced in suggested queries is likely to be associated to specific types of problematic suggestions.* Anecdotally, we know that stereotypical statements are likely to be about groups, while suggestions are more likely to be controversial due to the prefixes they are paired with. We expect to observe such intersections in our query data.

We believe these assumptions capture key aspects for tackling problematic query suggestions: Can we estimate the likelihood of a suggestion being perceived as problematic based on what the user typed? Are there suggestions more likely to be deemed problematic on their on, and thus should never be surfaced? Are certain types of problematic suggestions more likely to mention particular types of targets? If, e.g, the prefix plays a key role in how the suggestions are perceived, an initial decision on whether to trigger the query suggestion feature could be based on the prefix.

3.1 Experimental Framework Overview

Examining problematic query suggestions requires comprehensive collections of search scenarios where such suggestions might be surfaced. Thus, a first challenge is how to sample search scenarios in a way that it both surfaces sufficient examples of known types of problematic suggestions and casts a wide net to support further exploration of the problem space. To probe how various factors may lead to problematic suggestions, a second challenge is how to systematically manipulate them. Depicted in Figure 2, our study consists of five high-level steps:

Step 0: We determine a set of dimensions (and corresponding categories) to characterize problematic suggestions: context, content, and target (§3.2).

Step 1: We sample search queries from a large commercial search engine through a mix of random and keyword based sampling strategies that are grounded in our content and target typologies (*step 0*), in order to ensure sufficient support for known categories (§4.1).

Step 2: We generate synthetic data for our crowdsourcing experiments by sub-sampling prefixes, suffixes, and possible search intents from our initial search query samples (*step 1*), while controlling for factors such as the frequency of the prefix, suffix, and intent (§4.2).
Step 3: We deploy crowd experiments that elicit assessments of whether a suggestion is perceived as problematic, while systematically manipulating the prefix, suggestions, and search intent (§5.2).
Step 4: We map suggestions identified as problematic to categories from our content and target typologies (*step 0*) through additional crowdsourcing tasks (§5.3).
Step 5: We analyze dependencies between the perceptions of problematic and the commonality of search scenarios, their content and target, and contextual cues, among other factors (§6).

3.2 Problematic Query Suggestions: Definition & Dimensions

To study problematic suggestions, one has to first define what suggestions may be problematic. Given the exploratory purpose of our study, here we aim to cast a wide net and not be too prescriptive about what may or may not constitute a problematic suggestion.

Defining "problematic" suggestions. We use an inclusive working definition that incorporates a wide range of problematic suggestions mentioned in prior work e.g., [12, 15, 22, 64, 76, 87], and broadly consider problematic any suggestion that may be *unexpectedly offensive, discriminatory, biased, or embarrassing*, or may *promote deceit, misinformation or content that is in some other way harmful* (including adult or suicidal content). Such suggestions may reinforce stereotypes, or nudge users towards harmful or questionable patterns of behaviour. Third parties may also try to manipulate the suggestions to promote, for instance, a business or a website [82].

As with other latent constructs, even with this definition drawing a clear delineation between problematic and non-problematic cases may not be possible. To further ground our exploration, we assemble operationalizing typologies for each dimension we consider: *context*, *content*, and *target*.

D1. Problematic in context. While a query completion suggestion aims to capture a valid search intent and match the text a user started to type (query prefix), the same completion might be suggested to multiple query prefixes. While the prefix may influence how problematic a certain suggestion is perceived to be, some suggestions might be perceived as problematic for any query prefix. We thus distinguish between suggestions that are 1) problematic *regardless* of what the user typed (or on their own), and those that are 2) problematic *given* what the user typed.

To catalogue suggestions along the *content* and *target* dimensions, we turned to prior work on generating query suggestions [50, 63, 72], auditing suggestion engines [53, 66, 67], identifying problematic suggestions (e.g., offensive, partisan, adult [6, 15]), and examining their implications (e.g., discriminatory, defamatory, illicit [64, 71, 85]); including in-depth media articles covering anecdotes and issues with search query suggestions [24, 43, 81]. Then, following common practices in thematic analysis [1, 7, 61], we iteratively merged related concepts mentioned across the various threads of literature in a bottom-up fashion. To obtain manageable coding schemes—suitable for crowd annotations (§5)—we constructed coarse-grained categorizations that cover categories likely to be well represented in the data. We gather the remaining concepts into "catch-all" categories, further allowing us to also capture cases that might have been overlooked by prior work.

D2. Types of problematic suggestions. When users assess whether a suggestion is appropriate, the textual content and topic of the resulting query is often the most salient, providing cues on why suggestions might be perceived as problematic: are they offensive? do they reinforce stereotypes? could they nudge users towards harmful behavior? The resulting *content typology* (described in a greater level of detail in Table 5 in Appendix) includes the following categories:

- *Harmful speech:* might be construed as offensive or hateful (including promoting violence, intimidating, or reflecting derogatory attitudes towards individuals or groups).

- *Potentially illicit:* might be construed as condoning or constituting illicit speech (such as infringing on intellectual property) or as nudging users towards illicit activities.
- *Controversy, misinformation, & manipulation:* might be perceived as controversial or misleading, as nudging users towards conspiracy theories, or as manipulated to promote certain content.
- *Stereotypes and bias:* might be perceived as discriminatory towards certain groups (including racist, sexist, homophobic), or as endorsing certain ideological views.
- *Adult content:* contains racy terms, can nudge users towards pornographic or obscene content.
- *Other types of problematic content:* might be perceived as problematic on other grounds, including due to promoting animal cruelty or suicidal content, triggering memories of traumatic events, or relates to sensitive or emotionally charged topics.

D3. Targets of problematic suggestions. A typical building block in systems that support users in their search tasks is the recognition of entities like people or companies [33]. In fact, prior work has largely focused on problematic suggestions referencing entities [36, 54, 64], such as surfacing negative associations about individuals and companies, or reinforcing racial or gender stereotypes. We conjecture that suggested queries that reference certain entities—or more generally *targets* [62, 63]—may be more likely to be problematic. For example, hateful content is rather about groups or individuals than about say plants [23, 56, 85]. Our working *typology of targets* (detailed in Table 4 in Appendix), includes the following categories of targets:

- *Individuals:* the subject of the query is a public or private individual.
- *Groups:* the subject of the query is a group of individuals that share at least one characteristic (such as race, gender, age, occupation, appearance, disability, or country of origin).
- *Businesses:* the subject of the query is a specific business.
- *Organizations:* the subject of the query is an organization, institution, or agency, which can be governmental or non-governmental (but not a specific business).
- *Animals and objects:* the subject of the query is an animal, group of animals, or can be construed as an object or a group of objects.
- *Activities and ideas:* the subject of the query is a specific activity, action, or idea.
- *Other possible targets:* the subject of the query does not fit any other category.

4 DATA COLLECTION & GENERATION

Our data collection is shaped by two of our experimental objectives: 1) construct annotated collections of search query completion scenarios—operationalized as `<prefix><suffix><intent>` triplets—to help explore the effects of the query prefix, the user intent, or the query suggestion list on how suggestions are perceived; and 2) assess the prevalence and interplay of various types of problematic suggestions across the three dimensions of interest: *context*, *content*, and *target* (§3). There are, however, several practical challenges including: How many and what type of search scenarios (*queries*) to include in our experiments (and thus in these collections)? How to sample relevant search queries for each of these types? What other factors we should and could consider?

We adopt a two-stage approach to generate synthetic search query completion scenarios for our experiments: first, we sample candidate queries from a large corpus of search queries (§4.1), and then we sub-sample query prefixes, suffixes, and possible search intents from these queries (§4.2).

4.1 Sampling Search Scenarios

Past empirical evidence suggests that problematic suggestions may exhibit distinctive linguistic patterns [10, 82] and some search topics may be more prone to surfacing problematic suggestions [26, 53]. Our query sampling is thus centered on a key assumption: queries containing certain keyphrases have varying propensities of being seen as problematic. To vary the likelihood of surfacing problematic queries, we employ different heuristics to select key-phrases to match on search queries. To examine a wide range of suggestions, including across all scenarios covered by our typologies of problematic suggestions, we need to surface sufficient search scenarios for each type.

Query sampling. Thus, to also ensure enough variation in the type of problematic scenarios we are likely to observe, we exhaustively grounded our query samples in the (sub)category covered by our typology (see Table 4 and 5). The sets of key-phrases were assembled following standard practices in keyword-centered data collections [10, 28, 58], including phrases used in [10, 16] or specific to topics and categories mentioned in prior work (e.g., from lists of swear words [40], lists of lawmakers [89], or nationalities [69]), example phrases mentioned in the media [15, 24, 43], and ad campaigns [80, 85], among others. To allow matching on slight lexical and syntactic variations of each phrase (e.g., verb forms, plurals, different orders), we used these sets to define regular expressions while restricting the matching such that the phrases were included in full (or in part for longer phrases) in the query prefix. This also means that our samples are shaped by the limitations to keywords-based data collection [27, 57, 79]. The resulting samples, including a baseline, are:

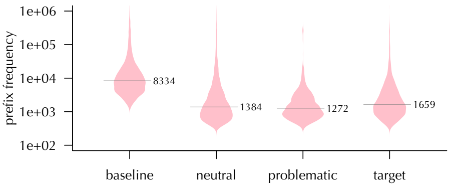
1. *Likely problematic* – queries that are more likely to be problematic than a random query. We sampled these queries using both problematic key-phrases (e.g., swear words, racy terms) and key-phrases related to a range of sensitive and vexing topics such as controversial (e.g., `abortion` [10, 55]), misinformation (e.g., `climate change` [55], `vaccination` [17, 86]), and hate speech (e.g., antisemitism [24], racism [22]), among others. To uncover a diversity of problematic scenarios, we assembled separate sets of key-phrases for each lower-level category in our content typology—see in Table 5 the (sub)categories along definitions and additional query and key-phrase examples.
2. *Contains a target* – queries concerning specific subjects or targets, such as those belonging to our categories of targets (§3.2, also listed in Table 4). To locate queries referencing a diversity of targets, we used phrases or names used to describe various types of targets. For instance, for individuals we searched for names of well known politicians (e.g., `Barack Obama`, `Donald Trump`), artists (e.g., `Lady Gaga`), scientists (e.g., `Jane Goodall`), and a variety of public figures (e.g., `Elon Musk`, `Bill Gates`); for groups we searched for terms or phrases used to refer to groups defined by gender (e.g., `girls`, `guys`), religion (e.g., `Muslims`, `Christians`), age (e.g., `young people`), occupation (e.g., `scientists`); and so on.
3. *Likely neutral* – queries with neutral prefixes that are expected to have a lower propensity of being problematic, as the prefixes do not relate to any sensitive or problematic topic. Following prior work [10, 62, 63], for this sample we used neutral key-phrases such as expressions indicating the user is explicitly asking a question or expressing an intent to do something. Specifically, we searched for queries starting with a question, including wh-questions, how, should, could, would (e.g., `could I try [swimming]`, `why do girls [hate math]`), or by expressing intent, including patterns like `intend|will|hope|expect to` (e.g., `I intend to [harm myself]`, `I hope to [kill my cat]`).
4. *Baseline sample* – queries satisfying a frequency threshold. Since the other samples may overlap with this base sample, to minimize the overlap in frequent queries among samples, for the base sample we also used query data from an additional 1-week period (from a different month).

We drew these samples from a large corpus of anonymized search queries from a major search engine (www.bing.com)—covering the entire month of February 2019—and extracted queries matching the key-phrases we selected for each sample. We restricted our samples to queries issued at least 5 times (15 for the base sample) during the observation period, and containing between 4 to 20 query terms—ensuring we can surface at least 1 term as suggestion and avoid suggestions that are too long. The resulting samples total over 20 million unique queries that were issued about 4 billion times by the search engine's users. Table 1 shows basic statistics for each of our samples.

Table 1. Dataset figures, including the number and frequency of queries, as well as their experimental viability.

Sample Type	Unique queries	Total frequency	Candidate prefixes	Viable prefixes	Period
Problematic	1.4M	46.6M	0.45M	606	02/2019
Target	6M	395.6M	1.7M	2,823	02/2019
Neutral	2.9M	75.3M	0.39M	3,351	02/2019
Baseline	19.3M	3,436M	2.75M	3,423	02-03/2019

(a) Summary statistics for each query sample. All prefixes are set to 3 terms, and experimentally viable prefixes have at least 45 unique suffixes in our samples.



(b) Prefix frequency distribution. Horizontal lines indicate the median frequency.

4.2 Generating Synthetic Search Completion Suggestion Scenarios

Illustrated in Figure 1, as users type the *prefix* of their query (informed by their *search intent*), search engines often make completion *suggestions*. Given that it is often not evident to users, the perception of these suggestions as problematic is expected to be agnostic to the mechanism used to generate them (e.g., based on popularity or other factors). For experimental purposes and without loss of generality, we assume that the suggestion engine relies on a corpus of past queries and their frequency to decide which query suffixes to select as candidate suggestions for a prefix [8, 50, 72].

To deploy crowd experiments where judges assess query completion suggestion scenarios, we transform our query samples into: 1) sets of distinct *prefixes*, and 2) lists of candidate *suggestions* and 3) user *intents* (or *final queries*). Using a factorial design (§5.2), for each prefix we select from our samples (4 query samples x 3 prefix selection heuristics x 80 prefixes), we test different search query completion scenarios by systematically varying both the possible search intents (2 query selection heuristics x 3 possible search intents) and the list of completion suggestions (3 suggestion lists x 6 completion suggestions x 2 suffix selection heuristics). Overall, we thus experiments with a total of about 200,000 `<prefix><suffix><intent>` triplets selected from our 4 query samples.

Prefix selection. Following prior findings that users are more likely to select suggestions at word boundaries, for longer prefixes, and after they typed about half of their query (e.g., [38, 44, 51]); for our experiments we only consider prefixes at word granularity as we split queries into prefix-suffix pairs and fix the prefix length to 3 terms—i.e., half of the highest average query length across our samples (6.15 terms in the neutral sample). For each prefix, we then extracted the list of available suffixes in our samples along their frequency. For each query sample we selected 3 sets of 80 prefixes each, according to prefixes' overall frequencies: 1) *common* prefixes—most frequent 80 query prefixes in each sample; 2) *rare* prefixes—least frequent 80 prefixes in each sample; and 3) *random* prefixes—a random set of 80 prefixes from each sample.

To test different search completion scenarios for a prefix, however, the inventory of candidate suffixes needs to be sufficiently large. To ensure this, for experimental purposes, we only use prefixes with over 45 unique suffixes, a threshold defined to ensure both no overlap between the 6 different suggestion lists (detailed below) and larger variations in the frequency of `<prefix><suffix>` pairs. Our prefixes are thus selected from a total of 8, 351 viable prefixes across all query samples.

Suggestions & user intent selection. Given a query prefix typed by a user, we assume the search engine shows 6 different completion suggestions—chosen given that the number of suggestions shown by major search engines varied over time between four to ten suggestions [47, 66, 67, 76]. Then, for each prefix we selected 6 different suggestion lists by sampling suffixes according to the frequency of the corresponding prefix-suffix pairs: (1) *common* suffixes—the 18 (3 suggestion lists x 6 suggestions) most common query suffixes for each prefix, and (2) *rare* suffixes—the 18 (3 suggestion lists x 6 suggestions) least common query suffixes for each prefix.

Similarly, to vary the possible *search intents* for each prefix, we selected and used the 3 most (respectively least) common `<prefix><suffix>` pairs. In our experiments, for each prefix we thus considered 6 distinct user intents, and for each prefix-intent pair 6 different suggestion lists.

5 CROWDSOURCED ASSESSMENTS

Through a mix of crowd assessment and categorization tasks (§5.1), we elicit input from crowd judges on 1) which query completion scenarios might be construed as problematic by search engine users (§5.2) and 2) to categorize such scenarios according to the *content* and *target* typologies (§5.3).

5.1 Crowd tasks characteristics

In our crowd experiments, we are interested in judges' subjective reactions rather than in gathering ground truth assessments. While content filtering policies may (and in some cases should) diverge from user assessments, to provide more nuance to our analysis we wanted to capture a range of perceptions that reflects differences in how people assess harms and risks [73, 74]. Thus, we purposefully avoid being too prescriptive about what makes a query completion suggestion problematic, and instead encouraged judges to think broadly about what may be perceived as biased, bothersome, or in some other way offensive by a search engine's users. Given that we may expose judges to a variety of problematic suggestions, we obtained IRB approval, and for all our crowdsourcing tasks we first obtained informed consent from our judges. Conforming to our IRB, we do not collect any demographic information from our judges.

In all our tasks, along our guiding definition of what may constitute a problematic suggestion (§3.2), the instructions also included a visual diagram highlighting the main elements of a search query completion scenario and the terminology we used throughout the task (similar to Figure 1). In addition, in the categorization tasks we also provided judges with detailed category definitions and several illustrative examples for each category.

Across all our tasks, we employed English speaking judges from North America, recruited from the Clickworker crowdsourcing platform (www.clickworker.com) via Microsoft's internal crowdsourcing tool—known as the Universal Human Relevance System (UHRS) [11, 29]—that allows easy task design and quality control. We capped the number of assessments per judge between 5% to 10% of the entire set of HITs per task (higher for the categorization tasks), paying on average about $15 (USD) per hour. For quality control, we restricted the tasks only to the top 30% judges on the platform; and to remove any remaining spammers and under-performing judges (and re-judge corresponding HITs), we used a mix of post-annotation and speed checks along qualification and spammer verification tasks (as applicable). Judges were paid for all HITs they performed, even when identified as spammers. In total, 364 trusted judges participated in our tasks.

5.2 Crowd Experiment Design

To manipulate several factors in order to discern their effect on how system provided suggestions for web search are perceived, we use a 3 (types of prefix samples) x 2 (types of suffix samples) x 2 (types of possible intents samples) factorial-like design [13, 19] to systematically vary the attributes of (§4.2): the *query prefix*, the possible *search intent*, and the *query completion suggestions*.

In the experimental setting, we showed judges a *query prefix* typed by a user, a list of six possible *query completion suggestions* and a *search intent* (a real search query starting with the prefix). We then asked them to separately assess each suggestion for whether it is 1) problematic *regardless of* what the user wrote, 2) problematic *given* what the user wrote, or 3) not problematic. While we encouraged judges to err on the side of marking suggestions as problematic when in doubt, we also asked them not to mark as problematic suggestions that are simply different from the user's search intent. Judges also answered a couple of questions intended to capture their views

Table 2. The distribution of crowd assessments aggregated at the `<prefix><suffix>` pair level. In the top table, the numbers reflect the distribution of problematic pairs, after the removal of those marked as false positives or as adult content. In the bottom table, the distribution also contains pairs marked with low confidence as adult content or as false positives (see text for details).

	Harmful	Illicit	Controversy	Stereotypes	Others
Problematic	15 (5.3%)	27 (9.5%)	73 (**25.7%**)	79 (27.8%)	90 (31.7%)
Target	10 (5.7%)	15 (8.6%)	40 (23%)	49 (**28.2%**)	60 (34.5%)
Neutral	13 (**7.0%**)	23 (12.4%)	42 (22.7%)	30 (16.2%)	77 (**41.6%**)
Baseline	8 (6.8%)	20 (**16.9%**)	23 (19.5%)	31 (26.3%)	36 (30.5%)

	Indiv.	Groups	Business	Org.	Animals/Obj.	Activ.	Others
Problematic	120 (11.8%)	535 (**52.8%**)	36 (3.6%)	15 (1.5%)	108 (10.7%)	53 (5.2%)	131 (12.9%)
Target	191 (**30.6%**)	181 (29.0%)	84 (**13.5%**)	28 (**4.5%**)	61 (9.8%)	19 (3.0%)	44 (7.0%)
Neutral	126 (18.4%)	91 (13.3%)	10 (1.5%)	11 (1.6%)	97 (14.2%)	121 (**17.7%**)	185 (**27.0%**)
Baseline	61 (14.8%)	93 (22.7%)	14 (3.4%)	10 (2.4%)	76 (**18.5%**)	36 (8.8%)	102 (24.9%)

on system-provided suggestions for web search: 1) Do you think *search engines should provide autocomplete suggestions* for the query prefix `<prefix>`? 2) Do you think *search engines should avoid surfacing problematic suggestions* for the query prefix `<prefix>`? Finally, they were provided with a free text box where they could leave additional comments on any concerns or thoughts they wanted to share with us. We collected two assessments for each `<prefix><suffix><intent>` triplet in our samples, resulting in a total of about half-million assessments that we overview in §6.1.

5.3 Categorizing Problematic Suggestions

We then annotated the `<prefix><suggestion>` pairs deemed problematic according to 1) their content and 2) who or what was mentioned. While in the experimental setting we asked judges to assess 6 suggestions at a time, in the categorization tasks we showed judges a single `<prefix> [<suffix>]` pair per HIT, collected between 3 (when all judges agreed) to 5 labels, and kept the majority label.

Types of Problematic Suggestions. To understand why some suggestions are construed as problematic, we used the categories in our content typology (§3.2) and for each problematic `<prefix><suggestion>` pair (§6.1) we asked judges to select the category that best characterizes the suggested query. In total, for this task we assessed 7,905 problematic `<prefix><suffix>` pairs (49.7% from problematic, 22.2% target, 15.4% neutral, and 12.7% baseline), collecting about 32,000 assessments. Across all samples, over 90% of problematic pairs were marked as either adult content (50%) or as a false positive (41.5%). Though carefully curated to help surface a variety of scenarios, the prevalence of adult queries in our samples is not surprising, but reflective of past observations on the pervasiveness of adult queries [34, 52]. These queries were not only prevalent in the likely problematic sample, but also in the target sample particularly for queries mentioning certain groups such as women or teens. From the remaining pairs, 263 (34.6%) were labeled as other types of problematic, 189 (24.8%) as stereotypes and bias, 178 (23.4%) as controversy, 85 (11.2%) as potentially illicit, and 46 (6%) as harmful speech. The breakdown for each sample is depicted in Table 2.

Targets of Problematic Suggestions. To further understand whether certain types of problematic suggestions are more likely to be associated with certain types of entities, e.g., person or organization, we also categorized problematic queries according to who or what was referenced by asking judges to select the best category from our target typology (§3.2). For this task, we filtered out `<prefix><suggestion>` pairs assessed either as adult content (often referencing groups or individuals) or as false positives with high agreement (>80%). As a result, we annotated 2,737 `<prefix><suffix>` pairs and collected about 11,000 assessments. Out of all pairs, 900 (32.9%) were found to be about groups, 498 (18.2%) about individuals, 462 (16.9%) about other types of targets, 342 (12.5%) about animals

Table 3. `<prefix> [<suffix>]` examples from our samples, with some paraphrased for clarity and anonymity.

`<prefix> [<suffix>]`	sample	prefix	suffix	Perceived as …
`what should we [do when we have flu]`	neutral	rare	rare	not problematic
`how long does [meth stay in your system]`	neutral	common	common	both problematic given and problematic regardless of the prefix, potentially illicit, and referencing an object
`alexandria ocasio cortez [swimsuit]`	target	common	common	problematic in context, as biased and stereotypical, and as referencing an individual
`neil degrasse tyson [arrested]`	target	common	common	problematic in context, as controversy, misinformation, and manipulation, and as referencing an individual
`blue cross blue [shield of ill]`	baseline	common	rare	problematic regardless of the prefix, as a false positive, and as referencing a business
`late term abortion [is never necessary]`	problem.	common	rare	problematic given the prefix, controversy, misinformation, and manipulation, and referencing other types of subjects
`little girl in [miniskirt]`	problem.	rare	common	both problematic given and problematic regardless of the prefix, adult content, and referencing a group

and objects, 229 (8.4%) about activities and ideas, 144 (5.3%) about businesses, and 64 (2.3%) about organizations. Table 2 contains the breakdown for each sample.

6 EXPLORATORY DATA ANALYSIS

We begin with a descriptive overview of the query completion scenarios assessments we collected (§6.1). We then contrast experimental results for rare versus common search scenarios to examine the possible effects of search voids [26] (§6.2), and explore associations between contextual factors and the prevalence of problematic scenarios (§6.3). Finally, we look at the interplay between our content and target typologies to see why some suggestions are deemed problematic (§6.4).

6.1 Crowdsourcing Assessments of Problematic Search Scenarios

Table 3 depicts a selection of crowd annotated `<prefix> [<suffix>]` pairs, illustrating the samples these pairs were drawn from, the heuristics used to construct the corresponding query completion scenarios, and the summary assessments we collected for each of them.

Crowd assessments aggregation and limitations. Given both our interest in capturing a range of perceptions and the subjective nature of our assessment task (where some judges might perceive something as problematic while others might not), our goal is *not* to determine some "ground truth" value for each `<prefix> [<suffix>]` pair. Judges may also lack the sensibilities, the cultural background, the knowledge or the experience to properly assess whether a given suggestion might be problematic. For instance, correctly assessing a scenario like `neil degrasse tyson [arrested]` (flagged only by 2 out of 12 judges) requires knowing this statement is false. Similarly, in the `blue cross blue [shield of ill]` scenario some judges might have thought of the abbreviation of Illinois state, "IL," instead of the "ill" term that refers to sick or incompetent (and may thus be problematic).

Many scenarios may thus not be salient enough for a majority of judges. Adopting a majority vote approach—otherwise a standard practice—to aggregating assessments may impose a too strict condition for what should be construed as problematic and may, as a result, inadvertently leave out scenarios that might rightly be perceived as inappropriate or harmful by only a minority of judges. For the remaining of our analysis, we consider a `<prefix> [<suffix>]` pair problematic if at least one judge marked it as such—following a *reporting model* used in applications where user complaint reports are reviewed irrespective of how many users submit them [14, 48].

While marking something as problematic if anyone flags it as such is prone to *false positives*, keeping a majority label is prone to higher *false negative* rates (and thus to overlooking problematic scenarios). To understand how this choice may affect our observations, Figure 3 shows the

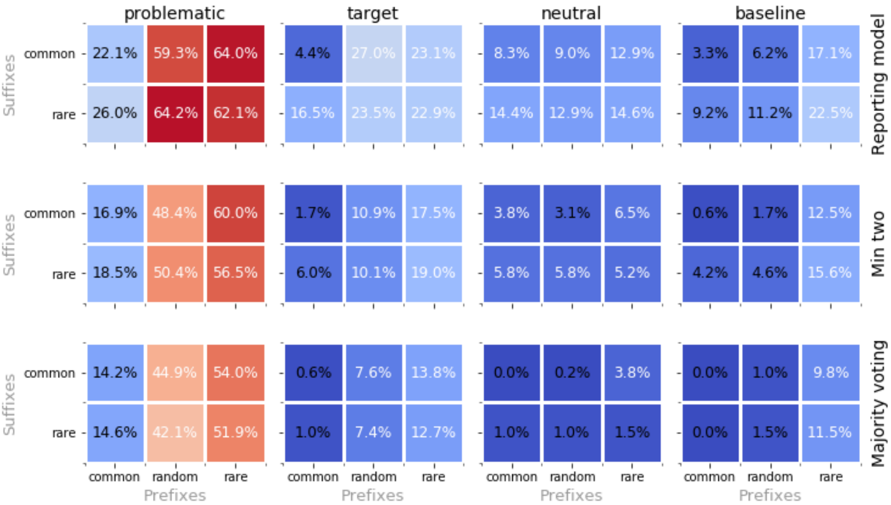


Fig. 3. Prevalence of problematic <prefix><suffix> pairs. The selection heuristic for prefixes and suffixes are shown on the X and Y axes, respectively. On the right side, the Y axis label indicates how the crowd assessments were aggregated to determine which of the <prefix><suffix> pairs were problematic.

prevalence of problematic <prefix> [<suffix>] pairs depending on how the assessments aggregation is done—including a hybrid heuristic ("min two") requiring two judges to flag a scenario as problematic in order to mark it as such. Across samples we see a gap of 8% to 20% gap in the prevalence of problematic scenarios, suggesting that a majority voting approach may overlook many problematic scenarios. While there are variations in the estimated prevalence, the overall patterns remain consistent (§6.2).

Problematic scenarios across query samples. We mixed query data sampling (§4.1) and selection (§4.2) heuristics to vary the prevalence and type of problematic scenarios we are likely to observe. Indeed, we see differences across our various experimental samples, reflected in both the prevalence and the type of problematic scenarios.

For instance, our likely-problematic sample helped surface almost four times more problematic <prefix> [<suffix>] pairs (49.9%) than our baseline sample (11.7%). Even the target and neutral (question or intent queries) samples led to higher rates of problematic suggestions than the baseline with 20.3% and 14.3%, respectively. There are also variations in the prevalence of different content and target categories across samples (Table 2). For example, the neutral sample seems to surface scenarios referencing activities (17.7%) at a higher rate than other samples (3.0%–8.8%); while queries on illicit activities appear more prevalent among problematic scenarios in the baseline sample (16.9%, after filtering false positives and adult queries) compared to other samples (8.6%–12.4%).

Exploring the "catch-all" categories. To account for lesser known or infrequent scenarios, we added a "catch-all" category to both the content and target typologies (§3.2). In our experiments, 263 problematic <prefix><suggestion> pairs for content and 462 for target were categorized as "Other types," respectively "Other targets." For content, these include known topics like animal cruelty (e.g., `how to kill [a bird]`) and suicidal thoughts (e.g., `what medication can [kill you]` or `can I just [die already]`), as well as a wider range of sensitive topics like traumatic events (e.g., `how to regain [purpose after bullying]`), relationship (e.g., `should i tell [my husband i cheated]`) and family issues (e.g., `when you don't [like your adult children]`). Similarly, for target, they include mentions of some legislation

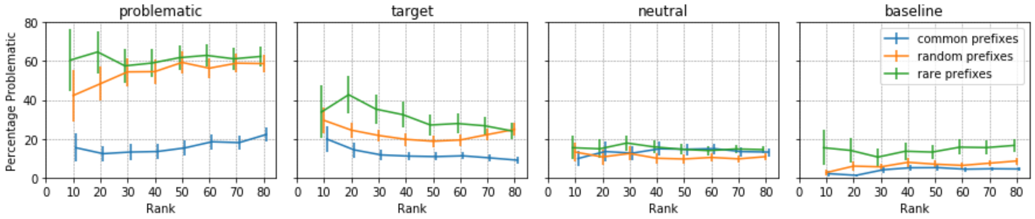


Fig. 4. Mean percentage of problematic suggestions for prefixes at a given rank by frequency (within different samples and prefix sets) when surfacing *common* suffixes. The bars depict the standard error of the mean. Similar trends were observed for rare suffixes.

(e.g., `death penalty in [png]`), emotions (e.g., `when you feel [worthless]`), body parts (e.g., `images of the [human body]`), or offensive behavior and symbolism (e.g., `what is so [wrong with blackface]`). Multiple target scenarios were also marked as "Other targets," e.g., `fox news breaking [hillary crimes]`.

6.2 Search Voids: Rare vs. Common Query Completion Scenarios

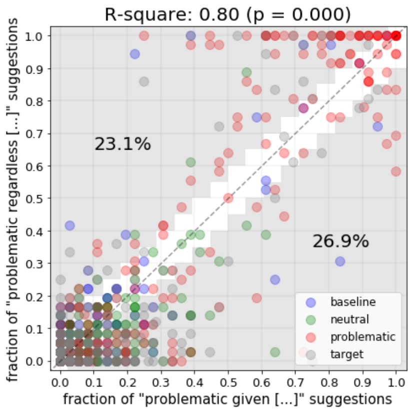
Golebiewski and boyd [26] refer to *data or search voids* as information needs with little or no relevant content in a corpus. Anecdotal evidence suggests that problematic search query suggestions are more likely to be surfaced in the absence of relevant queries in past logs [26, 50]. To explore this phenomenon, we operationalize data voids as *infrequent or rare prefixes and suffixes.* Figure 3 illustrates the prevalence of problematic scenarios when combining different heuristics for selecting prefixes and possible suggestions based on their historical popularity in our log samples. Overall, in our experiments, when we contrast common and rare prefixes, *rare prefixes* are more prone to problematic suggestions, in part due to there being fewer competing queries. When contrasting common and rare suffixes, overall *rare suffixes* also appear more likely to be problematic when surfaced as suggestions.

Problematic@Rank. Figure 4 further shows how the prevalence of problematic `<prefix><suffix>` pairs varies for prefixes at a given rank (according to their overall popularity) in different samples. We see similar trends as the curves for rare prefixes tend to dominate. We do not however remark any particular trends as a result of the relative rank of prefixes within our samples.

6.3 Problematic in Context

Here we explore associations between contextual factors (broadly construed) and the prevalence of problematic scenarios. We contrast cases where suggestions are problematic on their own (e.g., contain racy terms or slurs)—and more likely to be suppressed by current search engines—with cases where they are problematic due to the query prefix they were suggested for—e.g., the suggestion `teeth` is problematic in `women with no [teeth]`, but appropriate for `how to whiten [teeth]`.

Prefixes vary in their propensity to surface problematic suggestions. We start by aggregating crowd assessments at prefix level to examine differences in the prevalence of problematic suggestions for different prefixes, with judges flagging for some prefixes almost all suggestions from our samples as problematic (e.g., `70 year old [...]` or `why are black [...]`), while for other prefixes almost all suggestions were deemed appropriate (e.g., `images of the [...]` or `black churches in [...]`). Figure 5 shows example prefixes with different rates of problematic suggestions in our samples, while also depicting how some prefixes are not only more susceptible to problematic suggestions, but they also vary in their susceptibility to surfacing either suggestions that are problematic on their own or suggestions that are problematic due to the prefix they were paired with.



Type	Example prefixes		
	<=15%	<=50%	<=100%
Problematic	can a felon	how many abortions	women talk about
Target	women of color	alexandria ocasio cortez	boy forced to
Neutral	how to confirm	should i tell	how many school
Baseline	14 day forecast	i just want	very young girls

Fig. 5. Prefix distribution according to the fraction of suffixes seen as problematic given the prefix (x-axis) vs regardless of the prefix (y-axis). Each data point corresponds to a prefix. The prefixes that fall in the gray area have a higher or lower prevalence of suffixes being perceived as problematic in context by at least one judge—with the annotated percentages reflecting the fraction of prefixes in each of the two grayed areas, e.g., in our experiments 23.1% of prefixes were more prone to surface suffixes that are problematic on their own, while 26.9% were more prone to surface suffixes that were flagged as problematic due to the prefix they are suggested for. These numbers go to 30.1%, respectively 32.1%, when considering only rare suffixes. The table contains examples of prefixes with different rates of problematic suffixes in our samples. Best seen in color.

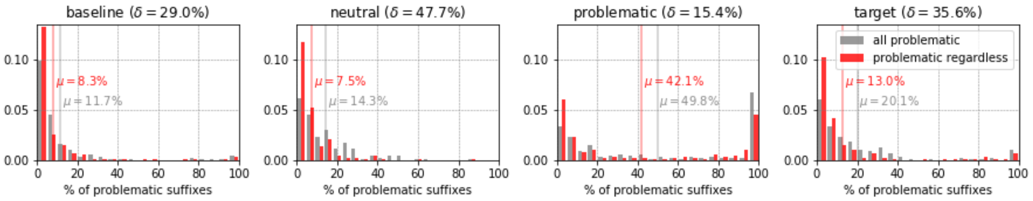


Fig. 6. Probability distribution of the prevalence of problematic suffixes when considering or ignoring those assessed as problematic given the prefix. δ represents the overall fraction of problematic `<prefix> [<suffix>]` pairs that would have been missed if only the suffixes assessed as problematic on their own would have been considered. μ is the mean % of problematic suggestions for the prefixes in each sample. Best seen in color.

Many suggestions can be problematic in context. This also raises concerns about potentially overlooking suggestions that are only contextually problematic at annotation time, if no distinction is made between the user input (*query prefix*) and what was suggested (*query suffix*). In fact, up to 26.1% of problematic `<prefix> [<suffix>]` pairs in our samples (35.6% for target, 47.7% for neutral, 15.4% for problematic, 29% for baseline) would have been missed if the suggestions were to be assessed on their own, and a fraction of those could be overlooked even when the assessment is done at the full query level—common practices when constructing training datasets and computationally filtering problematic suggestions [30, 45, 87, 88]. Figure 6 contrasts the distribution of the fraction of problematic suffixes per prefix when considering all scenarios assessed as problematic versus when considering only those scenarios where suggestions were problematic on their own. For the latter, we see the distributions are shifted towards lower ranges across all samples, misleadingly suggesting a significantly lower prevalence of problematic scenarios.

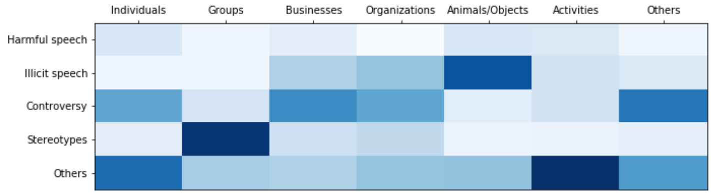


Fig. 7. The interplay between target and content categories for problematic scenarios. The color intensity shows the prevalence of content categories for each type of target. The values in the columns sum up to 100%.

The effect of search intent and judge stance on assessments. We also collected assessments while varying the possible search intents (§4.2), allowing us to examine whether knowing a user search intent may influence how judges assess query completion suggestions. For all our samples, we used paired t-tests to compare how the suggestions were assessed when judges were presented with either rare or common search intents and found no statistically significant differences ($p > 0.1$). While this suggests that knowing users' intents may have no impact on how suggestions are perceived, it may also be confounded by the search intents not being salient or distinct enough.

To examine whether the judges' views on system-generated suggestions influence their assessments, we also collected self-reports on whether—in the judge's opinion—search engines should provide query completion suggestions and whether problematic suggestions should be suppressed (§5.2). A majority of our judges consistently reported to believe that search engines should generally provide query completion suggestions (95%)—which we dub *pro-suggestions*—while only 49% consistently reported that problematic suggestions should be avoided—dubbed *pro-moderation*.[1] When checking if the *pro-suggestions* and *pro-moderation* judges assess query completion suggestions differently from the others, we found a relation between *pro-moderation* views and problematic judgements ($\chi^2 = 51.8, p < 0.01$)—who appear to be surprisingly slightly less likely to assess suggestions as problematic; but no difference for the *pro-suggestion* judges ($\chi^2 = 0.007, p = 0.93$).[2]

6.4 Query Types and Targets

We observed about 2 times more problematic `<prefix> [<suffix>]` pairs in our target sample compared to the baseline sample, suggesting queries containing targets might be more prone to be deemed problematic when surfaced as suggestions. Anecdotally, we know that queries of certain types or topics are more likely to be associated with certain types of targets (e.g., hate speech often targets groups and individuals)—but we do not know if this is generally the case. We examine this by looking at the interplay between the target and content typologies (§3.2).

Figure 7 shows how in our experiments target categories are more frequently related to particular types of content in our samples: each ⟨content, target⟩ cell estimates the probability that a *problematic* `<prefix> [<suffix>]` pair mentioning a certain target would match a specific content category across our experimental samples. Individuals are frequently the subject of other types of problematic scenarios (such as relationship issues `how to tell [him I cheated]` or `how to marry [your cousin]`) and controversy, misinformation, and manipulation (e.g., `what was wrong [with nancy pelosi mouth]` or `alexandria ocasio cortez [bankrupt]`). Problematic scenarios mentioning groups often reflected

[1] We consider a judge to be *pro-suggestions* if they have never reported to believe that suggestions should not be surfaced, and *pro-moderation* if they have always reported that problematic suggestions should be avoided.
[2] Note that in our experiments the search completion scenarios were randomized.

stereotypes and biased beliefs (e.g., `why are black [women so mad]` or `what a good [wife should do for her husband]`). When businesses and organizations were mentioned, the suggested queries were more likely about controversy, misinformation, and manipulation (e.g., `23 and me [scam]`) and potentially illicit activities (e.g., `microsoft word [torrent download]`). Queries referencing animals and objects were predominantly about potentially illicit activities (e.g., `how long does [meth stay in your system]`), while those referencing activities were predominantly categorized as other types of problematic scenarios (e.g., `just want to [die]`). Finally, scenarios referencing other types of targets were most often about controversy, misinformation, and manipulation (e.g., `how many abortions [are done at 9 months]`) and other types of problematic (e.g., `age of consent [in Thailand]`).

7 DISCUSSION

Query suggestion engines aim to assist users in completing their search tasks. While some suggestions exactly match a user search intent, others may be educational or even serendipitous, inspiring alternative searches and explorations. In fact, users' behavior may be "subtly influenced by exposure to query possibilities [they] may not have considered if left to [themselves]" [15]. Problematic query suggestions can thus have a variety of potentially harmful effects. Our study suggests a wide range of scenarios where suggestions are problematic, blind spots due to data annotation practices, and factors that make search engines more prone to surfacing problematic suggestions.

Accounting for Context

Our results show that contextual cues—including how common a query or an informational need is and what prefix a suggestion was surfaced for—may be associated with varying incidence of problematic scenarios, and with implications for the designers of both the suggestion engines and those of the frameworks to identify and suppress problematic cases.

Understanding the idiosyncrasies of search completion scenarios. While we have seen that certain prefixes are more prone to problematic suggestions than others, we do not know why the resulting queries appear more frequently in historic logs: do they reflect some prevalent needs or human nature, or are they the result of attempts to game the system by adversaries (e.g., [81, 82])? While prefixes like `girls [...]` or `women [...]` have millions of possible suffixes, we observed that the problematic ones tend to show up more frequently in past logs, making them more likely to be being picked up by suggestion engines relying on log data [8, 50]. For instance, while adult queries are known to represent a large fraction of web search queries [34, 52] and dominate the search space for these and other innocuous prefixes—being also more aggressively suppressed by search providers—we know less about why scenarios like `women [should be in the kitchen]` also tend to make it among popular suffixes.

To tease apart how various cues—either used to generate suggestions (e.g., the query prefix) or that may affect how suggestions are perceived (e.g., the judge stance)—relate to the observed prevalence of problematic scenarios, we sought to capture properties of the *query prefix*, *suffix*, *intent*, *type and target*, as well as the *judge stance* (§5.2). Yet, for all of these, we only explore a few (tractable) properties. For instance, for the *query prefix* we considered several types and popularity levels—operationalized through different data sampling (§4.1) and selection strategies (§4.2)—while fixing others like the prefix length. Future work should investigate additional user and informational cues, e.g., how the length of the prefix relates to the likelihood of surfacing problematic scenarios.

Operationalizing search intent and other user factors. The user intent in particular is hard to measure in production settings, and hard to operationalize and simulate in offline experiments. Examining the comments left by some of our judges (§5.2), we see that while some of them hinted at the search intent when justifying their assessments (e.g., "*[u]ser is looking for something related to* *[business] location. The results are okay, but they are mostly about [the business] headquarters*"), in our experiments, knowing the search intent does not appear to affect how judges assessed search query completion suggestions. However, the effect of running a query is often long lasting, as there is a correlation among not only the within sessions queries, but also among temporally distant queries [65]. The broader context within which the user issues a query may also affect how they assess the provided suggestions, and may thus capture better their search intent than a single query.

Accounting for personalized search suggestions. Problematic suggestions can also affect those issuing the search query if e.g., their dignity is compromised [49]. This is particularly true if the suggestions are assumed to be personalized—as additional `<prefix> [<suffix>]` pairs might be seen as problematic due to harms that arise from suggesting negative associations with the users themselves. This could include suggestions related to the interests, looks or the users' health and well-being. Suggesting `... [bed bugs by yourself]` when a user typed `how to kill [...]` may imply the system deduced the user has a bed bugs infestation issue. Other examples may include stereotypical or prejudicial associations like those highlighted by Sweeney [78] for online ads that associate someone's name with suggestive ads about arrests. For instance, suggesting `... [my arrest records]` when a user typed `how to find [...]` may hint the system assumed that such records may exist, as nothing that the user typed warranted such a suggestion.

Towards Preempting Problematic Suggestions

Research on fairness, accountability, and transparency in computational systems is building an ever-expanding set of tools for measuring and correcting biases and problematic scenarios that we know to look for [59]. However, techniques for preempting future issues that may not yet be on a product teams' or research community's radar are not nearly as well developed or understood. Addressing this gap requires studies like ours that deep dive into specific application areas.

Constructing training datasets of problematic scenarios. When a suggestion engine relies on trained models, the importance of context in recognizing problematic suggestions indicates that both modeling and data labeling assumptions need to be carefully selected. For data collection specifically, our results demonstrates that identifying and showing annotators the prefix and the suffix separately is necessary—otherwise, without knowing which part of a full query is the prefix and which the suggestion, the subtlety may be lost. In fact, failing to make this distinction may result in overlooking an important fraction of problematic `<prefix> [<suffix>]` pairs—in our experimental samples that ranged between 15% to 47% (§6.3). Combinations of various elements from our framework can be leveraged to surface a wider range of problematic scenarios and build more comprehensive training datasets. For instance, varying and mixing topically or semantically grounded sampling strategies (like those grounded in our content and target categories §4.1) with behavioral cues (like the popularity of a search query §4.2) can help manipulate the propensity of observing various types of problematic scenarios (§6.2–§6.4).

Biases in populations, behaviors, and logs. As in our study, the type of scenarios being surfaced is bounded by how the experimental or training datasets are sampled and generated, as well as by the quality of crowd annotations and self-reports [57]. Predictive text applications may also affect how users formulate their queries [2], and thus the characteristics of the resulting datasets. While auditing a specific search engine or query completion service was outside the scope of this study, understanding how representative the problematic scenarios we examined are of those surfaced by various services could help practitioners further prioritize specific types of problematic suggestions.

We also examined whether the judges' beliefs and values impact their assessments, observing that judges who self-reported to favour moderation spotted—to our surprise—less problematic scenarios. Additional work is needed to explore if this is an artefact of our study design or e.g., if judges that appear to favor moderation tend to apply a more narrow definition of what should be construed as problematic. Understanding whose viewpoints get amplified when certain query suffixes are presented as suggestions might offer further insights into how to avoid them by leveraging latent properties of the users issuing those queries. For our study we also only recruited English speaking crowd judges from North America, and did not collect demographic information. More diverse crowds may help us identify an even larger range of problematic suggestions, while more information about our judges may help understand differences in how people assess potential harms and why some scenarios are salient to only a minority of judges.

Changing the suggestion engines. Although our study assumes a frequency-based model for query completion, the results provide guidance for the design of alternative models, including those based on more sophisticated natural language generation (NLG) techniques. Since all these methods typically rely on training models on historical log data, which may contain an array of problematic examples (§6), alternative mechanisms that either recognize and deemphasize that content, or that place more emphasis on less problematic language data sources may be more desirable and should be considered. Even when data do not contain problematic examples, adversaries can exploit search or data voids to inject problematic content. Our results indicate that search voids do make the query suggestion engine more prone to problematic suggestions e.g., even up to 3 times more prone to surfacing problematic suggestions in the presence of rare query prefixes. Thus, leveraging rare search scenarios for training purposes may help preempting more problematic scenarios, particularly in the e.g., anticipation of new events [84]. Alternatively, algorithms should recognize topics or areas with weak support or low confidence before suggesting a suffix.

Conclusions

Exploratory in nature, our study is the first effort to tease apart some of the factors that may influence how problematic search query suggestions are perceived to be. Starting with a multi-dimensional inventory of problematic search suggestions grounded in prior literature, we took a mixed-methods approach primarily based on integrating a range of query sampling strategies with crowdsourced experiments and categorization studies. Contrary to assumptions underlying current data collection and blacklisting protocols for problematic queries, the context in which a suggestion is being made plays a critical role in whether the suggestion is considered problematic. Collectively, our findings deepen our understanding of how searchers may be impacted by predictive technology and how we could preempt problematic scenarios.

ACKNOWLEDGMENTS

We would like to thank Mohamed Musbah, T.J. Hazen, Michael Golebiewski, and Luke Stark for helpful discussions and support throughout the project, and to our anonymous reviewers for their detailed and constructive feedback. We are also grateful to Emery Fine, Adam Ferguson, and Susan Powers for feedback on the design of the initial pilots of our crowd experiments.

REFERENCES

[1] Mohammed Ibrahim Alhojailan. 2012. Thematic analysis: A critical review of its process and evaluation. *West East Journal of Social Sciences* 1, 1 (2012), 39–47.

[2] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 128–138.

[3] Paul Baker and Amanda Potts. 2013. 'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies* 10, 2 (2013), 187–204.

[4] Ziv Bar-Yossef and Naama Kraus. 2011. Context-Sensitive Query Auto-Completion. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. Association for Computing Machinery, New York, NY, USA, 107–116.

[5] Karrisa Bell. 2017. LinkedIn's new messaging feature makes it easier to clear your inbox. https://mashable.com/2017/10/24/linkedin-smart-replies-ai/

[6] Erik Borra and Ingmar Weber. 2012. Political Insights: Exploring partisanship in Web search queries. *First Monday* 17, 7 (2012).

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] Fei Cai, Maarten De Rijke, et al. 2016. A survey of query auto completion in information retrieval. *Foundations and Trends® in Information Retrieval* 10, 4 (2016), 273–363.

[9] Robyn Caplan. 2018. *Content or Context Moderation?* Technical Report. Data & Society, New York.

[10] Sergiu Chelaru, Ismail Sengor Altingovde, Stefan Siersdorfer, and Wolfgang Nejdl. 2013. Analyzing, detecting, and exploiting sentiment in web queries. *ACM Transactions on the Web (TWEB)* 8, 1 (2013), 6.

[11] Wei-Chu Chen, Siddharth Suri, and Mary L Gray. 2019. More Than Money: Correlation among Worker Demographics, Motivations, and Participation in Online Labor Market. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 134–145.

[12] Anne SY Cheung. 2015. Defaming by Suggestion: Searching for Search Engine Liability in the Autocomplete Era. *Comparative Perspectives on the Fundamentals of Freedom of Expression"(Andras Koltay, ed), Forthcoming* (2015).

[13] Linda M Collins, John J Dziak, and Runze Li. 2009. Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs. *Psychological methods* 14, 3 (2009), 202.

[14] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.

[15] Nicholas Diakopoulos. 2013. Sex, violence, and autocomplete algorithms. *Slate (Aug. 2, 2013)* (2013).

[16] Nick Diakopoulos. 2014. Algorithmic defamation: the case of the shameless autocomplete. *Tow Center for Digital Journalism* (2014).

[17] Renee Diresta. 2018. The Complexity of Simply Searching for Medical Advice. https://www.wired.com/story/the-complexity-of-simply-searching-for-medical-advice/.

[18] Shiri Dori-hacohen, Elad Yom-tov, and James Allan. 2015. Navigating Controversy as a Complex Search Task. In *Proceedings of the 1st International Workshop on Supporting Complex Search Tasks (SCST 2015).*

[19] Hermann Dülmer. 2007. Experimental plans in factorial surveys: random or quota design? *Sociological Methods & Research* 35, 3 (2007), 382–409.

[20] Carsten Eickhoff, Jacek Gwizdka, Claudia Hauff, and Jiyin He. 2017. Introduction to the special issue on search as learning. *Information Retrieval Journal* 20, 5 (Oct 2017), 399–402.

[21] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of within-Session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14).* Association for Computing Machinery, New York, NY, USA, 223–232.

[22] Steve Elers. 2014. Maori are scum, stupid, lazy: maori according to Google. *Te Kaharoa* 7, 1 (2014).

[23] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. In *Twelfth International AAAI Conference on Web and Social Media.*

[24] S. Gibbs. 2016. Google alters search autocomplete to remove 'are Jews evil' suggestion. https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-suggestion.

[25] Tarleton Gillespie. 2018. *Custodians of the Internet.* Yale University Press, New Haven and New York.

[26] Michael Golebiewski and danah boyd. 2018. Data Voids: Where Missing Data Can Easily Be Exploited. *Data & Society Research Institute* (2018).

[27] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Social Networks* 38 (2014), 16–27.

[28] Valentina Grasso and Alfonso Crisci. 2016. Codified hashtags for weather warning on Twitter: an Italian case study. *PLoS currents* 8 (2016).

[29] Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. The crowd is a collaborative network. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 134–147.

[30] Parth Gupta and Jose Santos. 2017. Learning to Classify Inappropriate Query-Completions. In *Advances in Information Retrieval*, Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (Eds.). Springer International Publishing, Cham, 548–554.

[31] Chris Hoffman. 2018. Bing Is Suggesting the Worst Things You Can Imagine. https://www.howtogeek.com/367878/bing-is-suggesting-the-worst-things-you-can-imagine/

[32] Desheng Hu, Shan Jiang, Ronald E Robertson, and Christo Wilson. 2019. Auditing the Partisanship of Google Search Snippets. *positions* 16, 57 (2019), 58.

[33] Alpa Jain and Marco Pennacchiotti. 2010. Open entity extraction from web search query logs. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 510–518.

[34] Maryam Kamvar, Melanie Kellar, Rajan Patel, and Ya Xu. 2009. Computers and Iphones and Mobile Phones, Oh My! A Logs-Based Comparison of Search Users on Different Devices. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. Association for Computing Machinery, New York, NY, USA, 801–810. https://doi.org/10.1145/1526709.1526817

[35] Anjuli Kannan, Peter Young, Vivek Ramavajjala, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, and Marina Ganea. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. ACM Press, San Francisco, California, USA, 955–964.

[36] Stavroula Karapapa and Maurizio Borghi. 2015. Search engine liability for autocomplete suggestions: personality, privacy and the power of the algorithm. *International Journal of Law and Information Technology* 23, 3 (2015), 261–289.

[37] Mark T Keane, Maeve O'Brien, and Barry Smyth. 2008. Are people biased in their use of search engines? *Commun. ACM* 51, 2 (2008), 49–52.

[38] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2013. User Model-Based Metrics for Offline Query Suggestion Evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 633–642.

[39] Ashiqur R. KhudaBukhsh, Paul N. Bennett, and Ryen W. White. 2015. Building Effective Query Classifiers: A Case Study in Self-Harm Intent Detection. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1735–1738.

[40] Chris Kirk. 2013. The Most Popular Swear Words on Facebook. http://www.slate.com/blogs/lexicon_valley/2013/09/11/top_swear_words_most_popular_curse_words_on_facebook.html.

[41] Rachel Kraus. 2018. Gmail SmartReply may be creepy, but they are catching on like wildfire. https://mashable.com/article/gmail-smart-reply-growth/

[42] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 417–432.

[43] Issie Lapowsky. 2018. Google Autocomplete Still Makes Vile Suggestions. https://www.wired.com/story/google-autocomplete-vile-suggestions/

[44] Yanen Li, Anlei Dong, Hongning Wang, Hongbo Deng, Yi Chang, and ChengXiang Zhai. 2014. A Two-Dimensional Click Model for Query Auto-Completion. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*. Association for Computing Machinery, New York, NY, USA, 455–464.

[45] Yuli Liu, Yiqun Liu, Ke Zhou, Min Zhang, Shaoping Ma, Yue Yin, and Hengliang Luo. 2016. Detecting promotion campaigns in query auto completion. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 125–134.

[46] A Mahdawi. 2013. Google's autocomplete spells out our darkest thoughts. *The Guardian* 22 (2013).

[47] Dan Marantz. 2013. A Look at Autosuggest. https://blogs.bing.com/search/2013/02/20/a-look-at-autosuggest/

[48] J Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. *Available at SSRN 2602018* (2015).

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

[50] Bhaskar Mitra and Nick Craswell. 2015. Query Auto-Completion for Rare Prefixes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1755–1758.

[51] Bhaskar Mitra, Milad Shokouhi, Filip Radlinski, and Katja Hofmann. 2014. On User Interactions with Query Auto-Completion. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*. Association for Computing Machinery, New York, NY, USA, 1055–1058.

[52] Vanessa Murdock, Charles LA Clarke, Jaap Kamps, and Jussi Karlgren. 2012. Report on the workshop on search and exploration of x-rated information (SEXI 2013). In *ACM SIGIR Forum*, Vol. 47. ACM New York, NY, USA, 31–37.

[53] Eni Mustafaraj, Emma Lurie, and Claire Devine. 2020. The Case for Voter-Centered Audits of Search Engines during Political Elections. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 559–569.

[54] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York City.

[55] Olivia Solon and Sam Levin. 2016. How Google's search algorithm spreads false information with a rightwing bias. https://www.theguardian.com/technology/2016/dec/16/google-autocomplete-rightwing-bias-algorithm-political-propaganda.

[56] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*.

[57] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2 (2019), 13.

[58] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In *Proceedings of Eighth International AAAI Conference on Weblogs and Social Media*.

[59] A Olteanu, J Garcia-Gathright, M de Rijke, MD Ekstrand, A Roegiest, A Lipani, A Beutel, A Lucic, A-A Stoica, A Das, et al. 2019. FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval. In *SIGIR Forum*, Vol. 53.

[60] Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-Scored Analysis of Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 370–386.

[61] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 994–1009.

[62] Marius Pasca. 2014. Acquisition of Noncontiguous Class Attributes from Web Search Queries. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 386–394.

[63] Marius Paşca. 2014. Queries as a Source of Lexicalized Commonsense Knowledge. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1081–1091.

[64] Kacy Popyer. 2016. Cache-22: The Fine Line between Information and Defamation in Google's Autocomplete Function. *Cardozo Arts & Ent. LJ* 34 (2016), 835.

[65] Matthew Richardson. 2008. Learning about the world through long-term query logs. *ACM Transactions on the Web (TWEB)* 2, 4 (2008), 21.

[66] Ronald E. Robertson, Shan Jiang, David Lazer, and Christo Wilson. 2019. Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 235–244.

[67] Ronald E. Robertson, David Lazer, and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 955–965.

[68] Alex Rosenblat, Tamara Kneese, et al. 2014. Algorithmic Accountability. *The Social, Cultural & Ethical Dimensions of "Big Data"* (2014).

[69] Benjamin Elisha Sawe. 2019. Largest Ethnic Groups And Nationalities In The United States. https://www.worldatlas.com/articles/largest-ethnic-groups-and-nationalities-in-the-united-states.html.

[70] Paul Sawers. 2018. Google rolls out smart replies to Hangouts Chat. https://venturebeat.com/2018/12/06/google-rolls-out-smart-replies-to-hangouts-chat/

[71] Barry Schwartz. 2015. Google Removes "How Can I Join ISIS" Autosuggestion. https://searchengineland.com/google-removes-can-join-isis-autosuggestion-214394

[72] Milad Shokouhi. 2013. Learning to personalize query auto-completion. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 103–112.

[73] Michael Warren Skirpan, Tom Yeh, and Casey Fiesler. 2018. What's at Stake: Characterizing Risk Perceptions of Emerging Technologies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 70.

[74] Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein. 1980. Facts and fears: Understanding perceived risk. In *Societal risk assessment*. Springer, 181–216.

[75] Michael L Smith. 2013. Search Engine Liability for Autocomplete Defamation: Combating the Power of Suggestion. *Journal of Law, Technology Policy* 2013 (Nov. 2013), 313–336.

[76] Danny Sullivan. 2011. How Google Instant's Autocomplete Suggestions Work. https://searchengineland.com/how-google-instant-autocomplete-suggestions-work-62592

[77] Danny Sullivan. 2018. How Google autocomplete works in Search. https://www.blog.google/products/search/how-google-autocomplete-works-search/

[78] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[79] Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[80] Marc van Gurp. 2013. Google's autocomplete feature is shocking too about black men. https://osocio.org/message/googles-autocomplete-feature-is-shocking-too-about-black-men/

[81] Jace Vernon. 2015. How YouTube Autosuggest and Google Autocomplete Can Work In Your Favor. https://marketinghy.com/2015/01/youtube-autosuggest-google-autocomplete-can-work-favor/

[82] Peng Wang, Xianghang Mi, Xiaojing Liao, XiaoFeng Wang, Kan Yuan, Feng Qian, and Raheem Beyah. 2018. Game of Missuggestions: Semantic Analysis of Search-Autocomplete Manipulations. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*.

[83] Ingmar Weber, Venkata Rama Kiran Garimella, and Erik Borra. 2012. Mining Web Query Logs to Analyze Political Issues. In *Proceedings of the 4th Annual ACM Web Science Conference (WebSci '12)*. Association for Computing Machinery, New York, NY, USA, 330–334.

[84] Stewart Whiting, Andrew James McMinn, and Joemon M. Jose. 2013. Exploring Real-Time Temporal Query Auto-Completion. In *Dutch-Belgian Information Retrieval (DIR) Workshop*.

[85] UN Women. 2013. UN Women ad series reveals widespread sexism. https://www.unwomen.org/en/news/stories/2013/10/women-should-ads

[86] Julia Carrie Wong. 2019. How Facebook and YouTube help spread anti-vaxxer propaganda. https://www.theguardian.com/media/2019/feb/01/facebook-youtube-anti-vaccination-misinformation-social-media.

[87] Harish Yenala, Manoj Chinnakotla, and Jay Goyal. 2017. Convolutional Bi-directional LSTM for Detecting Inappropriate Query Suggestions in Web Search. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 3–16.

[88] Harish Yenala, Ashish Jhanwar, Manoj K Chinnakotla, and Jay Goyal. 2018. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics* 6, 4 (2018), 273–286.

[89] Caitlin Yilek. 2017. The 9 most and least popular Capitol Hill politicians on Facebook. https://www.washingtonexaminer.com/the-9-most-and-least-popular-capitol-hill-politicians-on-facebook.

A CONTENT AND TARGET TYPOLOGIES

- Table 5 detailing the content typology, including examples and pointers to related work.
- Table 4 detailing the target typology, including examples and pointers to related work.

Received January 2020; revised June 2020; accepted July 2020

Category	Working definitions	Mentions in prior work	Example target (p/s: queries)
Individuals	References to a public or private person, who may or may not be explicitly named	actors [62, 63], politicians [6, 53, 83], congresspeople, public figures [16]	*ruth ginsburg* (p: ruth ginsburg [dead yet]) *my dad* (s: should I kill [my dad])
Groups	References to a group of individuals that share at least a common characteristic, such as race, gender, age, occupation, appearance, disability, or country of origin	teams [63]; ethnic and religious group [22, 24]; women [46]; protected groups, nationalities [76]; identity groups [3]; latinas, black, asian girls [54]	*muslims* (p: muslims try to [conquer through numbers]) *children with adhd* (s: how to punish [adhd child])
Businesses	References to a specific business	companies [16], companies, universities, newspapers [63]; school [12]	*Macy's* (p: macy's is [scamming shoppers]) *CNN* (s: should we punish [cnn]) *Starbucks* (s: should I boycott [starbucks])
Organizations	References to an organization, institution or agency, which can be governmental or non-governmental (but not a business); or a group of for-profit organizations if they are not specifically identified (e.g., news media instead of CNN, social media instead of Twitter)	companies [16], companies, universities, newspapers [63]; school [12]	*mainstream media* (p: mainstream media is [destroying america]) *UNICEF* (p: UNICEF is running [a scam]) *travel companies* (s: don't waste money on [travel companies])
Animals & objects	References to an animal, a group of animals, or anything that may be construed as an object or a group of objects.	flowers, food, cars, drugs [63], animals [15]	*cat* (s: how to poison [a cat]) *knife* (p: how to use a knife [to kill])
Activities & ideas	References to a specific activity, action, or idea	actions [60]; voting [53]	*cutting yourself* (p: cutting yourself is [stupid]) *crying* (p: crying is [emotional blackmail])
Other targets	References to concepts like ideologies, religions, programs, health issues, a situation someone may find themselves in, or other types that do not fit other categories	diseases, disorders, medical treatments [18, 62]; religions [76]; elections [53]; weather phenomena [84]	*bipolar disorder* (p: bipolar disorder is [fraud]) *vaccination* (p: vaccination [herd mentality]) *science* (p: science should [stay out of faith])
Generic, no target	There is no identifiable target or subject	...	(what [the heck]) (damn damn [damn])

Table 4. Typology of query targets (with definitions) and their relationship to related aspects mentioned in prior work. It also contains examples targets (and corresponding queries mentioning them) for each category.

Category	Working definitions (and sub-categories)	Mentions in prior work	Keywords (p/s: query)
Harmful speech	*Hate speech*: suggestions that could be perceived as hateful or that intend to intimidate or promote violence, against a group or its members. *Intimidates & promotes violence*: suggestions that may steer users towards acting violently or that aim to intimidate certain individuals. *Offensive speech*: suggestions that dehumanize, insult, or ridicule, actively seeking to embarrass or harm reputation. *Discriminatory speech*: suggestions showing known or existing bias, prejudice, or intolerance, perpetuating, employing negative stereotypes, or encouraging feelings of fear or disgust towards a group or individual. *Defamation & derogatory speech*: suggestions that defame someone by suggesting negative associations, including suggestions of dishonesty or involvement in illicit activities. *Profane language*: suggestions including any sort of slurs, expletives, swear or curse words.	extreme violence, cause harm to others, rude or discourteous or exhibiting lack of respect [87]; violence [15]; racism [22]; hate speech, violence related suggestions [76]; negative connotations [64]; antisemitic, sexist and racist [24]	*punch* (p: should i punch [my mother]) *hit* (s: should women be [hit by men]) *deported* (s: arabs should be [deported]) *poison* (p: which poison can kill [an adult fast])
Potentially illicit	*Facilitates illicit activities*: suggestions condoning & constituting illicit speech, infringing on intellectual property, copyright rights or trademark agreements, or that facilitate or nudge users towards illicit activities. *Privacy breaching*: suggestions revealing unwanted details from someone's past or anything that may be construed as sensitive or personal information. *Terrorist or extremist propaganda*: suggestions that may steer or help users find extremist content related to terrorist or extremist activities like recruiting or sponsoring. *Defamation & derogatory speech*: See above. *Child abuse & pornography*: suggestions related to child abuse or child pornography.	related to an activity which is illegal [87]; defamation [16]; illicit sex [15]; piracy related, personally identifiable information in suggestions [76]	*heroin* (s: trustworthy website to [buy heroin]) *fake passports* (s: how to get [fake passports]) *beat child* (s: how to [beat my child])
Controversy, Misinformation, and Manipulation	*Controversial topics*: suggestions that seem to endorse one side of a known controversial debate. *Misinfo., disinfo. or misleading content*: suggestions that promote information that is factually incorrect, or that reinforce or nudge users towards conspiracy theories. *Coordinated attacks & suggestions manipulation*: suggestions that occur as a result of attempts to manipulate the search or suggestions results, such as by promoting certain businesses or by trying to affect someone's reputation.	issue-related queries [6]; false information, organized attacks [49]; rumors and conspiracies [16, 53]; opinionated queries, controversial topics [10, 18]; partisan cues [32]; misinformation [55]; controversies, manufactured suggestions, fake queries [76]	*hoax* (s: climate change is [a hoax]) *staged* (s: 911 was [staged]) *vaccines* (p: vaccines are [dangerous]) *divorce lawyer* (p: divorce lawyer [nashville LAW_FIRM_NAME])
Stereotypes & Bias	*Ideological bias*: suggestions that validate or endorse views that belong to certain ideological groups, or that promote stereotypical beliefs about an ideological group. *Systemically biased suggestions*: suggestions about certain topics that are systematically biased towards a group, reinforcing sensitive associations between the group & negative attributes or stereotypical beliefs. *Discriminatory speech*: See above. *Defamation & derogatory speech*: See above. *Offensive speech*: See above.	gender bias, partisan, politically charged queries [6]; perpetuating stereotypes, make negative associations with an individual [49]; politically bias [42, 53, 67]; reinforcing and perpetuating stereotypes [3, 54]; defamatory or libelous suggestions, harmful or negative associations [68]; defamation [12]	*refugees* (p: refugees are [taking jobs]) *women* (p: women need [to dress modestly]) *girl* (s: running like [a girl]) *black men* (p: black men [are lazy])
Adult queries	*Adult content*: suggestions that contain pornography-related terms or steer users towards pornographic/obscene content. *Child abuse*: See above.	sex-related [16]; child pornography [15]; porn and adult-content related [34, 76]	*naked* (p: naked girls [videos])
Other types	*Animal cruelty*: suggestions that may steer users towards information about how to harm animals. *Self-harm and suicidal content*: suggestions that may steers someone towards hurting themselves. *Sensitive topics*: suggestions that may trigger memories of traumatic events or be considered sensitive or emotionally charged by certain groups due to historic or cultural reasons.	causing harm to oneself [87]; animal cruelty [15]; self-harm [39]	*strangle dog* (p&s: how to strangle [a dog]) *hitler* (p: hitler is [my god]) *hurt myself* (s: I want to [hurt myself])

Table 5. Typology of problematic suggestions (with definitions) and their relation to related aspects mentioned in prior work. It also contains examples of keywords we used to identify queries likely to belong to these categories of problematic suggestions, along with example of problematic queries containing these keywords. Sub-categories repeat (e.g., a query can both contain harmful speech and be potentially illicit).