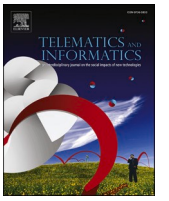
Telematics and Informatics 85 (2023) 102068

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele



Trapped in the search box: An examination of algorithmic bias in search engine autocomplete predictions

Cong Lin [a,1], Yuxin Gao [c,1], Na Ta [a,*], Kaiyu Li [b], Hongyao Fu [a]

[a] School of Journalism and Communication, Renmin University of China, China
[b] School of Information Technology, York University, Canada
[c] University of Toronto, Toronto, Canada

ARTICLE INFO

Keywords:
Search engine
Autocomplete
Algorithmic bias
Mediatization
Digital inequality
Algorithm auditing

ABSTRACT

This paper examines the autocomplete algorithmic bias of leading search engines against three sensitive attributes: gender, race, and sexual orientation. By simulating search query prefixes and calling search engine APIs, 106,896 autocomplete predictions were collected, and their semantic toxicity scores as measures of negative algorithmic bias were computed based on machine learning models. Results indicate that search engine autocomplete algorithmic bias is overall consistent with long-standing societal discrimination. Historically disadvantaged groups such as the female, the Black, and the homosexual suffer higher levels of negative algorithmic bias. Moreover, the degree of algorithmic bias varies across topic categories. Implications about the search engine mediatization, mechanisms and consequences of autocomplete algorithmic bias are discussed.

1. Introduction

Search engines are crucial infrastructures in the process of information seeking and retrieval. Nevertheless, factors ranging from user characteristics to system functionalities can shape people's search behaviors during the process (Azzopardi, 2021). "The search engine manipulation effect" (Epstein and Robertson, 2015: E4512) shows that biased search rankings driven by algorithms can shift the political voting preferences of undecided voters by 20 % or more. Taking the searching process as a whole, the influence in reality occurs to users earlier than the presentation of result pages. In a typical searching scenario, regardless of whether the user has an exact query, as he/she types in the search box, algorithm-based query predictions appear right below the user's input and change with each keystroke even before the user clicks "enter" (Yew, 2019). This autocomplete feature (also referred to as autosuggestion) is claimed to be designed to optimize and accelerate the search process, which helps reduce typing by about 25 percent on average (Sullivan, 2018).

Despite existing advantages, the potential pitfalls of autocomplete are not neglectable. Since predictions are determined based on users' previous activities and the popularity of queries (Wang et al., 2018), autocomplete could be a mirror of society, reflecting people's attitudes and biases with billions of user searches per day. In 2013, a UN Women ad campaign found demeaning predictions from Google, such as "(woman should) stay at home" (Mahdawi, 2013). Noble (2018) declared search engines "algorithms of oppression" by reporting many offensive and immoral completions that present a high level of racism, especially for Black women. All these denouncements indicate the existence of algorithmic bias, which concerns the ethical risks inherent in the operation of

* Corresponding author at: No.59, Zhongguancun Street, Haidian District, Beijing.
E-mail address: tanayun@ruc.edu.cn (N. Ta).
[1] Cong Lin and Yuxin Gao contributed equally to this work and shared the first authorship.

https://doi.org/10.1016/j.tele.2023.102068
Received 7 June 2023; Received in revised form 5 October 2023; Accepted 26 October 2023
Available online 29 October 2023

algorithms and has gained momentum in recent years. According to Kordzadeh and Ghasemaghaei (2022), algorithmic bias is often referred to as the phenomenon that algorithmic processes produce discriminatory results that violate the norms of justice and equality and adversely impact particular people or communities in the workplace or society when used to automate or assist decision-making. Discriminatory autocomplete predictions produced by search engines represent a typical form of algorithmic bias calling for more urgent attention, especially since they present in the seemingly facilitative, easy-to-use digital tools that people take for granted.

Nevertheless, although an increasing number of media reports and commentary studies in recent years have criticized the problematic autocomplete predictions (e.g., Elers, 2014; Lapowsky, 2018; Miller and Record, 2017), much of the discourse has been framed through examples of typical cases that reflect social bias. However, Graham (2023) argues that the stereotyping embedded in search engines should be termed "aggregated discrimination" since it might not be clear on an individual level but is overwhelmingly discriminatory when evaluating on a larger sample. It is therefore essential to quantitatively inspect the nuanced group-specific discrimination based on large-scale autocomplete results data.

Moreover, algorithmic bias not only has its technical facet involving the manifestations of biases in algorithms' outcomes, but also its social aspects pointing to existing human biases affecting certain underprivileged and marginalized communities (Favaretto et al., 2019). This implies that algorithmic bias should be discussed in the context of social inequality. Researchers have proposed that digital inequality research should look more into big data-related questions such as inequality and the negative effects of algorithmic decision-making for vulnerable population groups (Lutz, 2019). However, currently few studies have inspected and discussed the association between autocomplete algorithmic bias and the digital inequality.

Responding to these gaps, the present study intends to employ the method of algorithm auditing (Sandvig et al., 2014) to examine the output of the search engine autocomplete algorithm. We highlight the problem of algorithmic bias of search engine autocomplete predictions (SEAPs), and refer to user search queries as "query prefixes" (Hazen et al., 2020; Kato et al., 2013). Specifically, we first simulate the query prefixes to collect autocomplete predictions, and then operationally measure the built-in algorithmic bias to explore the disparities in algorithmic outputs with respect to demographic attributes, namely, gender, race, and sexual orientation. Overall, this paper aims to further foster an in-depth understanding of algorithmic bias and its affinity with digital inequality.

2. Literature review

2.1. Algorithmic bias in the information age

Emerging biases brought about by algorithmic technologies have been reported in various contexts (e.g., Angwin et al., 2016; Lambrecht and Tucker, 2019). Although there is no universally accepted definition of algorithmic bias, researchers have referred to relevant manifestations of algorithmic bias from multiple disciplines. Computer scientists point out that in the process of algorithm decision-making, fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics. Thus, they are scrutinizing algorithms whose decisions are skewed toward a particular group of people (Mehrabi et al., 2021). From the perspective of social science, researchers are more concerned with the inheritance of algorithmic bias to social discrimination and its impact on social equity (Bigman et al., 2022). Hajian and Domingo-Ferrer (2012) distinguish between two forms of direct and indirect discrimination in data mining. Direct discrimination refers to algorithmic procedures that discriminate against minorities or disadvantaged groups on the basis of sensitive attributes related to group membership such as race, gender, or sexual orientation, while indirect discrimination means procedures that might intentionally or accidentally discriminate against a minority without explicitly mentioning discriminatory attributes. In this paper, we follow the logic of direct discrimination and define algorithmic bias as discrimination by algorithm outputs against individuals or groups based on sensitive attributes.

Examining the internal mechanism, the common refrain "Bias In, Bias Out" captures the essence of algorithmic bias constraints (Mayson, 2018). Friedman and Nissenbaum (1996) make an early contribution to the study of potential biases in computer systems and classify them into three categories: preexisting, technical, and emergent, referring to the bias roots in social practices and institutions, the technical constraints or considerations, and the emergent biases in the context of human use, respectively. Subsequent studies of algorithmic bias mechanisms mostly refer to the influence of the feedback loop, i.e., the potential accumulation of bias in the main components of algorithmic data production, algorithmic programming, and algorithm-human interaction (Chen et al., 2020; Mehrabi et al., 2021). In essence, algorithms trained based on data produced by biased decision-makers or generated from social prejudices against particular groups will inevitably inherit bias (Rambachan and Roth, 2020).

Regarding the autocomplete algorithm, although SEAPs predicting people's queries are generated by machine learning (ML) systems trained on digital traces of many individuals' collective search behaviors, their correctness and usefulness cannot be guaranteed (Miller and Record, 2017). With training data of the model left unchecked, queries expressing violence, hatred, racism, pornography, and profanity can be mined and learned by the algorithm and thus cause offensive or harmful bias in query predictions (Hazen et al., 2020). This paper aims to investigate algorithmic biases, particularly negative discrimination, exhibited in large-scale search engine autocomplete predictions.

2.2. Mediatization, algorithms and digital inequality

In an era when media technologies are deeply involved in and influence our lives, the interaction between media saturation and the corresponding socio-cultural change animated by media is of great concern in mediatization research (Couldry and Hepp, 2017; Hepp, 2013). According to (Lundby, 2009: 24), mediatization describes the process whereby communication refers to media and uses media so that media in the long run increasingly become relevant for the social construction of everyday life, society and culture as a whole.

Andersen (2018) discusses how search engines, algorithms and databases shape our everyday communicative actions as they make us internalize, and act along the modes of searching, ordering and archiving. Simply relying on keyword-based searches and calculating logic-based algorithms could make us implicitly manipulated by certain power structures. For example, a habitual searching behavior leaves us oblivious to the lure of the commercial forces behind search engines (Mager, 2012), while algorithms are increasingly recognized as active agents in shaping what kind of knowledge is organized, given access to and our perceptions of what counts as knowledge (Gillespie, 2014).

Given the role of mediatization in shaping social culture and realities, there is a need to be wary of the potential negative consequences of media technologies. Disentangling and narrowing social inequalities has consistently been an essential academic concern, and the mediatization of society also begs questions of how the media environment is formative regarding reducing or amplifying social inequalities Hepp and Hasebrink (2014: 15-17). Givskov and Deuze (2018) argue that our mediatizing societies required rich, nuanced investigation specifically with reference to the ways disadvantaged people experience inequality and manifest social diversity.

In recent years, researchers in the field of inequality and digital divides have extended their vision beyond technology access, skills and use to differentiated outcomes, whether they are positive or negative (Eubanks, 2018). Researchers have also proposed that digital inequality studies must include digital traces, algorithmic surveillance, and data-based discrimination in their "syllabus" because big data-related inequalities and the negative effects of algorithmic decision-making for vulnerable population groups have been obvious (Lutz, 2019; Micheli et al., 2018). In addition, digital inequality 3.0 identifies "well-being and life course" as one of the key directions of future research (Robinson et al., 2020) since big data technologies may lead to problems such as marginalization and stigmatization (Lupton, 2015; Nielsen et al., 2017). However, inadequate attention has been paid to psychological influences, for instance, subjective social well-being (Büchi et al., 2018). As an unexpected source, if SEAPs display unfairness and discrimination against specific individuals or social groups, it might directly lead to psychological impacts for the target population. This paper, guided by the theoretical perspective of mediatization, intends to critically examine the problematic outputs and consequences of algorithm-driven search engine autocomplete features, to reveal the group-specific disparities and discuss their relationship with digital inequality.

2.3. Search engine autocomplete algorithmic bias

Due to difficulties in data acquisition and research methods, comprehensive studies and verifiable evidence-based conclusions of autocomplete algorithmic bias are scarce. In recent years, some researchers have begun to empirically explore algorithmic bias by manually collecting autocomplete data or using tools to access search logs (Hazen et al., 2020). For example, a productive way was proposed to collect autocomplete results containing clear stereotypes using terms referring to particular identity groups (e.g., Black, White, gay) paired with question forms, e.g., 'what', 'which', and 'should' (Baker and Potts, 2013), which has been accepted and applied by numerous other studies (Al-Abbas et al., 2020; Roy and Ayalon, 2020). This approach has inspired the collection of the SEAP data for the current study.

For the measurement of algorithmic bias, existing studies mainly focus on the relative proportion of the positive or negative topic differences affiliated with social groups, especially with the help of the manual coding approach (e.g., Baker and Potts, 2013; Roy and Ayalon, 2020). This approach is not scalable for large-scale data analysis, and fails to measure the degree of bias, which is especially valuable given the possibility of algorithms yielding the same proportion of biased content for all groups. This paper makes an attempt to measure the algorithmic bias in the degree of discrimination drawing on multiple negative attributes of problematic autocomplete predictions.

Regarding personal demographic information (e.g., race) as protected or sensitive attributes, i.e., attributes that are mostly not legally allowed to be used as the basis for decisions (Beutel et al., 2019; Deho et al., 2022), studies have set their sights on the misrepresentation of different social groups. Prior studies have mainly focused on bias related to gender and race, while a few have looked at age, sexual orientation, and religion. For the gender attribute, existing research on autocomplete algorithmic bias has shown that men are more likely to be connected with positive words than women (Al-Abbas et al., 2020; Bonart et al., 2019). For the race attribute, the typical case would be Google's apology in 2015 for an algorithm that automatically tagged and labeled two African-Americans as "gorillas" after an innocuous online word search (Kasperkevic, 2015). These illustrate the importance of studying racial bias in search engines, especially since there is currently little quantitative research on racial bias in the context of search engine autocomplete. Meanwhile, prejudice regarding sexual orientation also needs to be discussed because discrimination against homosexuals or gays is not only widespread in our society but also embedded in many databases used for algorithmic learning (Dixon et al., 2018; Zhang et al., 2020).

Given the above discussion, this paper focuses on the built-in algorithmic bias of SEAPs, especially toward three sensitive attributes contained in user search queries: gender, race, and sexual orientation. It is worth noting that although these three attributes encompass broader categories, such as transgender in terms of gender, and bisexuality in terms of sexual orientation, this paper focuses mainly on the most dominant social categories reflected in user search query histories. Specifically, we raise the following research hypotheses:

H1. For gender-sensitive query prefixes, SEAPs for female-related query prefixes present higher algorithmic bias than SEAPs for male-related query prefixes.

H2. For race-sensitive query prefixes, SEAPs for query prefixes concerning (a) Black people and (b) Asian people present higher algorithmic bias than SEAPs for White people-related query prefixes.

H3. For sexual orientation-sensitive query prefixes, SEAPs for homosexuality-related query prefixes present higher algorithmic bias than SEAPs for heterosexuality-related query prefixes.

The degree of algorithmic bias might be closely associated with the context and topic of the discourse. For instance, gender stereotypes are prevalent in our society, but when we discuss gender inequality, we usually cite examples in specific areas or topics such as jobs and income, political participation, and the distribution of unpaid domestic and care work (Wijnhoven and van Haren, 2021). Previous research regarding search engine autocomplete algorithms have devoted to summarizing the differences in topic categories between subgroups (Baker and Potts, 2013; Roy and Ayalon, 2020). For example, Bonart et al. (2019) clustered three categories of topics based on autocomplete results of gender-specific politicians, i.e., political and economic, private and emotional, and local information. It indicates that topic categories are valuable when discussing group-specific biases because of their possible interaction with subgroups and sensitive attributes, which further motivates us to raise the following research questions:

RQ1. Will the query prefix topic category moderate the degree of algorithmic bias in SEAPs for different sensitive attributes?

3. Methods

In this paper, we employed the method of algorithm auditing through the logic of reverse engineering (Sandvig et al., 2014). Since the code embedded in algorithms remains black boxed, reverse engineering provides researchers with glimpses of how an algorithm works in practice by examining what data are fed into an algorithm and what output is produced (Kitchin, 2019). Studies on algorithmic bias have widely used this algorithm auditing method (e.g., Eslami et al., 2017; Otterbacher et al., 2018).

Following the logic of reverse engineering, we developed a research framework from control of data input, data acquisition to measurement of data output to explore search engine autocomplete algorithmic bias. By incorporating a variety of group- and topic-specific related terms, user query prefixes were simulated to collect autocomplete predictions from three leading search engines, namely, Google, Bing, and Baidu. To reflect the algorithmic bias, toxicity scores of the autocomplete predictions were computed using Google's Perspective API (Hosseini et al., 2017), which is trained from online corpora such as Wikipedia and the New York Times. We then analyzed and presented the discrepancies among different subgroups and topics.

3.1. Dataset: Query prefixes composition and autocomplete predictions acquisition

Each query prefix (*pf*) can be decomposed into two parts: a set of selection conditions (*sel*) and a semantic content (*con*), denoted as *pf* = (*sel, con*) or (*con, sel*). Each *sel* corresponds to one value of an assignment of sensitive attributes. In this paper, the sensitive attributes set is {gender, race, sexual orientation}. The values of gender, race and sexual orientation are {"male", "female"}, {"White", "Asian", "Black"}, and {"heterosexual", "homosexual"} respectively.

The design of *con* is challenging because of the following reasons. First, the *con* for each *pf* should be neutral so that the SEAPs could exactly reflect the bias from the sensitive attributes themselves. Second, the *con* should also be bias-sensitive so that we can collect SEAPs with clear stereotypes rather than irrelevant and generalized results. Therefore, we firstly used interrogative words in order to formulate *pf* in question form, an approach that can efficiently elicit stereotyping SEAPs (Baker and Potts, 2013). Moreover, we explored the algorithmic bias in terms of different topics by using eight stereotype-related concepts drawn from existing studies (Behm-Morawitz and Mastro, 2009; Dennhag et al., 2019; Shih et al., 1999), namely Appearance & physical, Identity, Behavior, Personality, Cognitive, Career, Domestic, and Relationship. The Appearance & physical refers to group biases related to ones' appearance, attire, and physical ability. The Identity, Behavior, Personality, Cognitive refer to group biases based on ones' social identity, actions and behaviors, character and personality cues, and rationality and cognitive skills, respectively. The Career, Domestic, and Relationship refer to group biases related to ones' roles in professions, household chores and childrearing, and closed relationships, respectively. Apart from above eight topics, the aforementioned interrogative words were classified into the nineth topic Others.

The construction of *pf* including *sel* and *con* follows several steps. For *sel*, we created a wealth of synonyms that refer to specific groups. For example, in the case of the female group the synonyms included "women", "girls", "wives", etc. For *con*, interrogative words such as "why", "when", "should" were firstly created drawing on previous literature (Baker and Potts, 2013; Roy and Ayalon, 2020). Then by referring to bias-sensitive queries from the study of Krieg et al. (2022), three authors manually constructed topic-related declarative words. For example, queries of appearance point to people's looks and body characteristics, we constructed similar words such as "hair", "figure", "face" in the topic of "Appearance & physical" accordingly. After words combination and deduplication, we discussed the incongruencies and settled comprehension divergence and finally kept neutral and bias-sensitive words. Considering the potential impact of the primary language, identical English and Chinese versions of *sel* and *con* were both constructed. In total, 2,292 *query prefixes* were generated. Examples include "why gay" with "why" as *con* and "gay" as *sel*, and "male athlete" with "athlete" as *con* and "male" as *sel*. Details of query prefixes construction can be found in Appendix A.

For the data collection, the two popular platforms Google and Bing as well as the most prominent Chinese search engine Baidu were considered as the data source. Google, Bing, and Baidu APIs were called to generate autocomplete predictions for the composed query prefixes. Given a *pf*, on the basis of collective searching trends on the platform, the search engine would automatically compute the probability of each potential prediction concerning each user's query history, location, query time, and collected personal preferences (Tahery and Farzi, 2020). Considering this, we employed computer programs to call the APIs directly, which can avoid the problems caused by historical search records and personalization. In addition, we set the locations to China and Canada to obtain the SEAPs in Chinese and English, respectively. In the data collection process, apart from all the constructed query prefixes, we further expand each *pf* by adding a space and one of 26 letters to the end, which enabled the retrieval of a much richer number of SEAPs. We finally obtained 274,450 SEAPs (46,384 from Google, 125,691 from Bing, and 102,375 from Baidu). Google and Bing returned fewer autocomplete results for Chinese prefixes, while Baidu provided more Chinese autocomplete predictions. After deduplication and data cleaning, 106,896 autocomplete results were used in further analysis (Detailed data distribution shown in Table 1).

Table 1
The distribution of collected SEAPs from three search engines.

		Google		Bing		Baidu	
Variable	Type	N	%	N	%	N	%
Syntactic feature	declarative	9503	30.57 %	13,040	41.95 %	8542	27.48 %
	interrogative	6931	9.14 %	25,801	34.03 %	43,079	56.82 %
Gender	male	6173	15.26 %	15,077	37.27 %	19,199	47.46 %
	female	4929	14.06 %	13,252	37.81 %	16,870	48.13 %
Race	White	804	5.41 %	5389	36.29 %	8658	58.30 %
	Asian	702	5.59 %	4975	39.60 %	6887	54.82 %
	Black	1001	6.01 %	5887	35.37 %	9754	58.61 %
Sex	heterosexual	1003	25.71 %	2297	58.88 %	601	15.41 %
	homosexual	4134	19.65 %	7397	35.16 %	9509	45.19 %
Language	Chinese	0	-	0	-	51,621	100 %
	English	16,434	29.73 %	38,841	70.27 %	0	-

3.2. Measurements of outcome variable: Negative bias

Since not all forms of bias manifest themselves in unfair or otherwise unethical acts, what this paper seeks to discuss is the negative bias that refers to unfair, problematic and discriminatory beliefs or expressions that one directs toward a particular individual or group. Researchers have endeavored to clarify what is biased autocomplete prediction (Miller and Record, 2017; Olteanu et al., 2020). Through a series of crowd-sourced experiments, an array of scenarios where autocomplete predictions are considered problematic surfaced (e.g., harmful speech, illicit activities, misinformation, stereotype, and adult query content). Benefiting from the success of natural languages processing models such as BERT and its variants (Sun et al., 2019), it is more practical to classify text by training a machine learning model on a big corpus with labels or fine-tuning a pre-trained model with few labeled data.

In this paper, we used the Perspective API from Google, a free service that uses machine learning models to score the perceived impact a short sentence might have on. The model of Perspective is trained on millions of comments from a variety of sources, including comments from online forums such as Wikipedia and the New York Times, across a range of languages including Chinese and English. For each SEAP, we computed scores including TOXICITY (rude, disrespectful, or unreasonable content), SEVERE TOXICITY (hateful, aggressive, or disrespectful content), IDENTITY ATTACK (hateful comments targeting someone because of their identity), PROFANITY (swear words or curse words), INSULT (insulting, inflammatory, or negative words), and SEXUALLY EXPLICIT (contains references to sexual acts or body parts). These scores are offered by the Perspective API as probabilities (from 0 to 1), indicating how likely it is that a reader would perceive the text content as containing the given attribute. To illustrate specifically how Perspective API scores different SEAPs in terms of the negative attributes above, we provided several examples from the datasets with more details in Appendix B.

To measure the algorithmic bias of SEAPs, all six toxicity-related scores were used for further analysis and we also composed an overall negative bias score for each prediction by using principal component analysis (PCA) for the dimensionality reduction of the above six indicators. We suppose the overall score constructed from multiple negative attributes, which reflect discrimination or defamation against individuals or groups in a given text, can be used as a measure of negative bias of autocomplete algorithms from a holistic perspective.

The overall negative bias score was deprived by firstly testing the Kaiser-Meyer-Olkin (KMO) of the six indicators. The higher the KMO, the stronger the communality between the indicators. It is generally acknowledged that a KMO value above 0. 6 is acceptable for principal component analysis. The result of KMO is 0.742 ($p < 0.001$) which means indicators involved are suitable for PCA. SPSS (version 26.0) was used to conduct the PCA and the eigenvalue-based extraction method was selected. According to the criterion of eigenvalue (4.802) greater than 1 and the cumulative contribution of the explained variance (80.031 %) which presented an ideal effect, a principal component was extracted. We obtained and normalized the weights of the six indicators by using the eigenvalue and loadings of the component matrix, which were then integrated into an overall score of negative bias. The negative bias score ranges from 0 to 1, with higher scores indicating more toxic text. Details of the KMO and PCA are provided in Appendix C.

3.3. The independent variables and control variables

Four independent variables that may influence the algorithmic bias were examined: gender, race, sexual orientation, and topic category, of which the first three are well-adopted sensitive attributes in algorithmic bias studies. Each SEAP was labeled a value from one of the three sensitive attributes: gender {“male”, “female”}, race {“White”, “Asian”, “Black”}, and sexual orientation {“heterosexual”, “homosexual”} according to corresponding labels of the prefixes used. For the topic category, as mentioned above, nine categories including“Appearance & physical”, “Identity”, “Behavior”, “Personal”, “Cognitive”, “Career”, “Domestic”, “Relationship”, and “Others” were applied to label each SEAP.

In addition, we identified several potential influencing variables, including text length, syntactic features and platforms. For text length, longer prefixes are more specific and less likely to elicit many autocomplete results. Meanwhile, it was noted different *pf* forms can elicit autocomplete predictions differently, so we took the syntactic feature (e.g., declarative or interrogative sentences) of the *pf* into consideration. Besides, previous studies have recommended further examinations of queries in different languages and social

environments to find if they can yield similar stereotypical representations of minority groups (Baker and Potts, 2013). Therefore, we conducted a comparison study by taking into account Google and Bing as search engine representatives in English environment and also Baidu in Chinese environment. We finally used platform source as a grouping variable, and text length and syntactic feature as control variables.

3.4. Data analysis

To test the research hypotheses, we performed analysis of covariates (ANCOVA) by using SPSS with negative bias score and six toxicity-related scores as dependent variables. The effects of gender, race and sexual orientation were examined with text length and syntactic feature as covariates. To detect whether sensitive attributes interact with topic categories, gender, race, and sexual orientation were compared on different topic categories using two-way analyses of variance (ANOVAs). On the premise that each sensitive attribute interacts with the topic category, we analyzed the simple effects of each sensitive attribute one by one. Effect sizes comparing both groups at each topic category were computed.

4. Results

4.1. Autocomplete algorithmic bias of sensitive attributes

Grouping by different search engine platforms, we present the results of the negative bias and toxicity-related scores with different sensitive attributes in Figs. 1-3. For gender, ANCOVA results of the overall negative bias indicated there was a significant difference between the degree of bias towards male and female on either Google ($F(1, 11098) = 1037.031$, $p < 0.001$, partial $\eta^2 = 0.085$), Bing ($F(1, 28325) = 2743.713$, $p < 0.001$, partial $\eta^2 = 0.043$), or Baidu ($F(1, 36065) = 1262.5$, $p < 0.001$, partial $\eta^2 = 0.071$). Fig. 1 shows that female-related prefixes are more likely to suffer higher degree of negative bias in SEAPs than men in all platforms. ANCOVA for the other six toxicity-related scores have also demonstrated the same significant male–female discrepancies (see Appendix D and Fig. 1) and verified the inequality of algorithmic bias in relation to gender on such three search engines. Thus, H1 is supported.

For race, the results of ANCOVA indicated that the overall negative bias was significantly different depending on the race type on Google ($F(2, 2502) = 500.863$, $p < 0.001$, partial $\eta^2 = 0.286$), Bing ($F(2, 16246) = 2416.385$, $p < 0.001$, partial $\eta^2 = 0.229$), and Baidu ($F(2, 25294) = 1720.388$, $p < 0.001$, partial $\eta^2 = 0.120$). Fig. 2 shows that compared to White people, Black people experienced higher levels of overall negative bias in all platforms. The results of group comparisons further indicated the significant difference on Google ($t(2502) = 22.218$, $p < 0.001$, $d = 1.056$), Bing ($t(16246) = 40.758$, $p < 0.001$, $d = 0.769$) and Baidu ($t(25294) = 49.493$, $p < 0.001$, $d = 0.731$). Thus, H2a is supported. Compared to White people, Asian people experienced lower levels of overall negative bias in

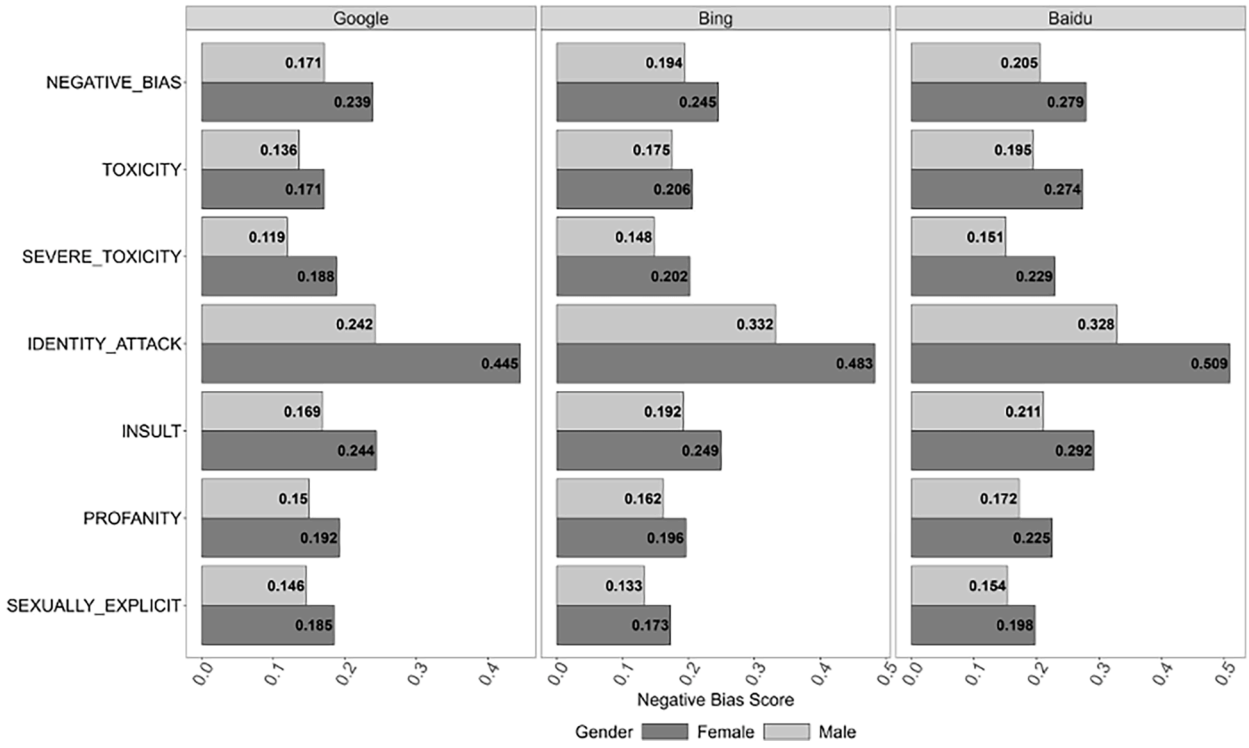


Fig. 1. Negative bias and toxicity-related scores for different genders and platforms.

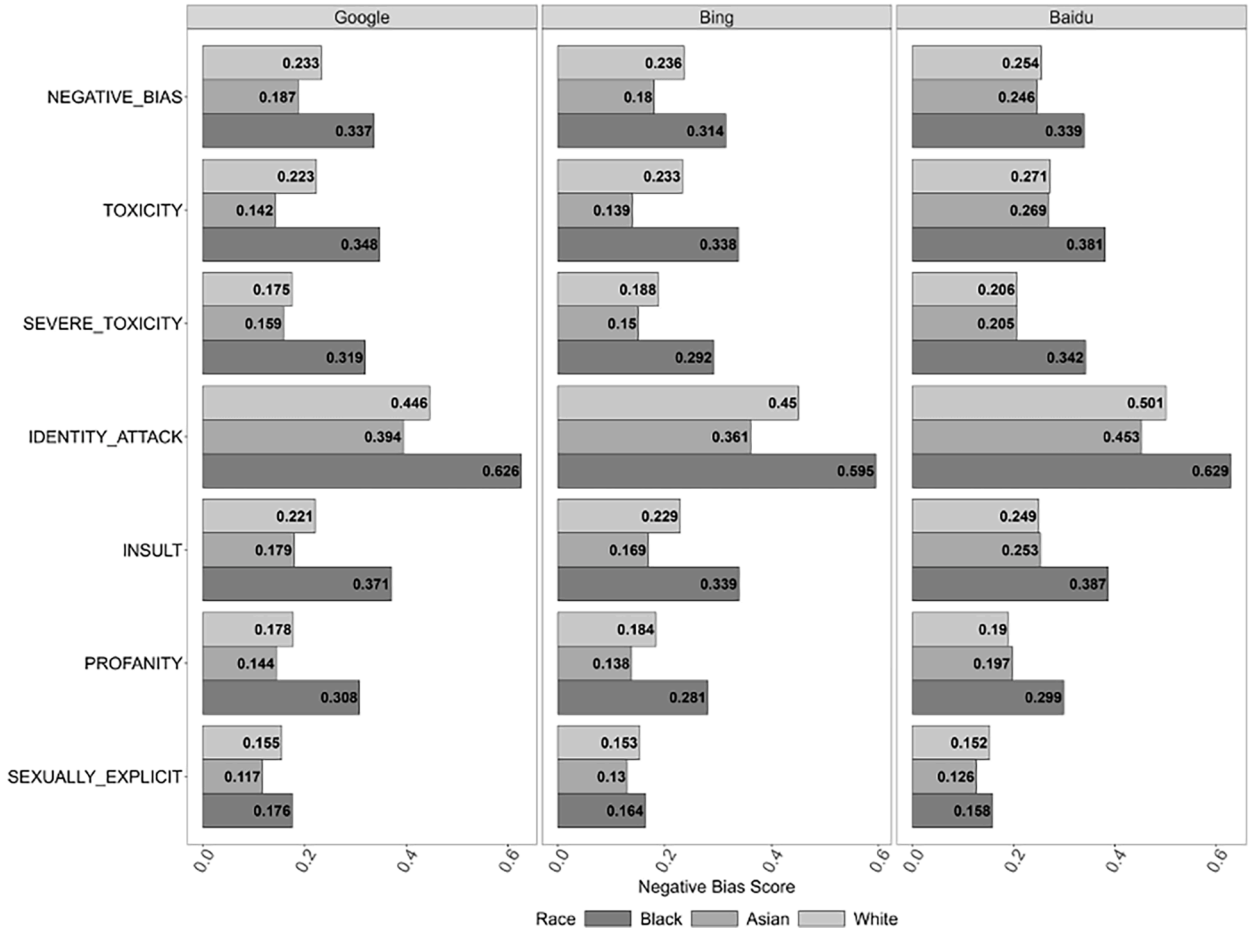


Fig. 2. Negative bias and toxicity-related scores for different races and platforms.

all platforms according to toxicity scores in Fig. 2. The results of group comparisons further validated this on Google ($t(2502) = -8.287$, $p < 0.001$, $d = -0.434$), Bing ($t(16246) = -28.4$, $p < 0.001$, $d = -0.559$) and Baidu ($t(25294) = 3.984$, $p < 0.001$, $d = -0.065$). Thus, H2b is not supported. ANCOVA for the other six toxicity-related scores have also demonstrated the similar pattens of discrepancies between White, Asian and Black people (see Appendix D and Fig. 2).

For sexual orientation, ANCOVA results of the overall negative bias indicated there was a significant difference between the degree of bias towards heterosexuality and homosexuality on Google ($F(1, 5133) = 2642.502$, $p < 0.001$, partial $\eta^2 = 0.340$), Bing ($F(1, 9690) = 2404.616$, $p < 0.001$, partial $\eta^2 = 0.199$), or Baidu ($F(1, 10106) = 8.468$, $p < 0.001$, partial $\eta^2 = 0.001$). Fig. 3 shows that, compared to heterosexuality, homosexuality suffered higher degree of algorithmic bias in all platforms. In general, ANCOVA for the other six toxicity-related scores on Google, Bing and Baidu have also demonstrated the same significant discrepancies between heterosexuality and homosexuality (see Appendix D and Fig. 3). Thus, H3 is supported.

In addition to discrepancies of subgroups, we can observe from Figs. 1-3 that among six toxicity-related scores, the two most prominent which presented high degree of algorithmic bias were “IDENTITY ATTACK” and “INSULT”, with “TOXICITY” being relatively severe for the race subgroup and “PROFANITY” for the sexual orientation subgroup as well. Meanwhile, among search engine platforms, Figs. 1-3 signify that the overall negative bias and specific toxicity-related scores were all relatively higher on Baidu than Google and Bing for gender and race, while for sexual orientation the scores were relatively lower on Baidu than on Google and Bing.

4.2. *Interaction effects between sensitive attribute and topic category*

Using the overall negative bias score as dependent variable, the results of the two-way ANOVA of topic category and each sensitive attribute suggest that the main effects of category for gender ($F(8, 75482) = 100.762$, $p < 0.001$, partial $\eta2 = 0.011$), race ($F(8, 44030) = 406.605$, $p < 0.001$, partial $\eta2 = 0.069$) and sexual orientation ($F(3, 13652) = 123.059$, $p < 0.001$, partial $\eta2 = 0.026$). The interaction effect of topic category and gender ($F(8, 75482) = 58.807$, $p < 0.001$, partial $\eta^2 = 0.006$), topic category and race ($F(16, 44030) = 37.515$, $p < 0.001$, partial $\eta^2 = 0.013$), topic category and sexual orientation ($F(3, 13652) = 157.105$, $p < 0.01$, partial $\eta^2 = 0.033$) were all significant. According to Table 2, the mean values of the negative bias score are higher for the female than the male, for

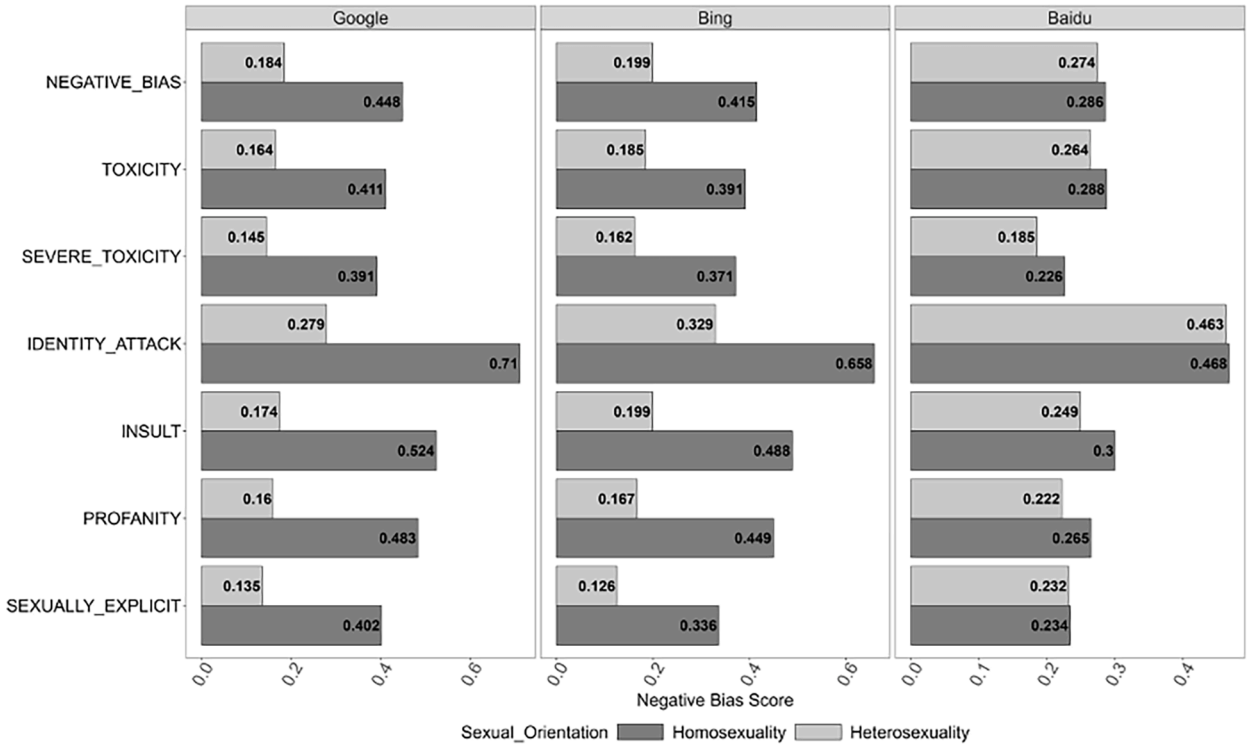


Fig. 3. Negative bias and toxicity-related scores for different sexual orientations and platforms.

Table 2
The mean values and standard deviation results of Two-way ANOVA.

	Gender		Race			Sexual Orientation	
	Mean (S.D.)		Mean (S.D.)			Mean (S.D.)	
	Male	Female	White	Asian	Black	Hetero.	Homo.
Appearance & Physical	0.211 (0.113)	0.279 (0.109)	0.259 (0.111)	0.241 (0.101)	0.352 (0.122)	0.299 (0.117)	0.4 (0.173)
Identity	0.195 (0.126)	0.284 (0.123)	0.278 (0.093)	0.265 (0.117)	0.376 (0.096)	0.163 (0.122)	0.408 (0.18)
Behavior	0.174 (0.098)	0.271 (0.105)	0.276 (0.081)	0.287 (0.088)	0.363 (0.106)	–	–
Personal	0.188 (0.118)	0.244 (0.11)	0.258 (0.093)	0.234 (0.101)	0.331 (0.116)	–	–
Cognitive	0.215 (0.118)	0.266 (0.106)	0.251 (0.106)	0.259 (0.089)	0.337 (0.113)	0.235 (0.111)	0.338 (0.192)
Career	0.193 (0.124)	0.242 (0.115)	0.222(0.114)	0.17 (0.092)	0.306 (0.119)	–	–
Domestic	0.206 (0.115)	0.254 (0.107)	0.221 (0.11)	0.158 (0.09)	0.313 (0.114)	0.141 (0.103)	0.296 (0.184)
Relationship	0.174 (0.128)	0.25 (0.121)	0.263 (0.11)	0.212 (0.094)	0.296 (0.128)	–	–
Others	0.172 (0.108)	0.286 (0.102)	0.274 (0.082)	0.301 (0.083)	0.372 (0.095)	–	–

the Black people than the Asian and White people, and for the homosexual than the heterosexual, in all different topic categories.

We conducted a follow-up test comprising an analysis of the simple main effect of topic categories and each sensitive attribute (see Table 3). The results indicated that for each topic category, the degree of bias in SEAPs varies significantly between query prefixes concerning people of different genders, races, and sexual orientations. The categories carrying severe bias also differed slightly between sensitive attributes. The partial η^2 value of 0.01 denotes a small effect; 0.06 denotes a medium effect; and 0.14 denotes a large effect. With this criterion, we can conclude that the topic categories with significant male-female disparity in the gender attribute include four types (behavior, identity, appearance & physical, and career), with behavior ($F(1, 75482) = 1455.88$, $p < 0.001$, partial $\eta^2 = 0.019$), appearance & physical ($F(1, 75482) = 1166.582$, $p < 0.001$, partial $\eta^2 = 0.015$) being the most prominent. For the race

Table 3
Results of the simple main effects for each sensitive attribute.

	Gender		Race		Sexual orientation	
	F	Partial η^2	F	Partial η^2	F	Partial η^2
Appearance& Physical	1166.582***	0.015	742.765***	0.033	219.82***	0.016
Identity	1147.88***	0.015	348.112***	0.016	2615.45***	0.161
Behavior	1455.88***	0.019	237.855***	0.011	-	-
Personal	578.915***	0.008	401.008***	0.018	-	-
Cognitive	430.809***	0.006	326.857***	0.015	289.982***	0.021
Career	767.01***	0.01	1909.996***	0.08	-	-
Domestic	271.43***	0.004	870.324***	0.038	96.288***	0.007
Relationship	331.464***	0.004	80.343***	0.004	-	-
Others	322.487***	0.004	40.065***	0.002	-	-

Notes: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$; Partial η^2, effect size.

attribute, the corresponding categories are more diverse including 7 types, and the most typical ones are career ($F(2, 44030) = 1909.996$, $p < 0.001$, partial $\eta^2 = 0.08$), domestic ($F(2, 44030) = 870.324$, $p < 0.001$, partial $\eta^2 = 0.038$) and appearance & physical ($F(2, 44030) = 742.765$, $p < 0.001$, partial $\eta^2 = 0.033$). For sexual orientation, the topic categories with the most significant discrepancy are identity ($F(1, 13652) = 2615.45$, $p < 0.001$, partial $\eta^2 = 0.161$). Therefore, RQ1 is answered.

5. Discussion and conclusion

Algorithm-driven search engines are deeply integrated into society as digital infrastructures. Although scholars have paid much attention to the algorithmic bias of search engine results pages, inadequate endeavors have been made to examine the autocomplete feature. With large-scale SEAPs elicited from leading search engines, this paper constructs metrics measuring the negative algorithmic bias in the degree of discrimination and reveals that autocomplete feature can yield socially-biased outcomes between social groups at the semantic level, thereby at the risk of compounding inequality in society. In addition, the topic categories discussed in SEAPs may moderate the bias degree and thus we investigate the interactive effects of stereotype-related topics and sensitive attributes.

5.1. Manifestations of search engine autocomplete algorithmic bias

The key findings of this paper are the manifestations of negative algorithmic bias around three sensitive attributes: gender, race, and sexual orientation. Specifically, we found query prefixes containing female terms are more likely to yield autocomplete results of a higher degree of negative bias than male-related terms in all platforms. For race, across all platforms, Black people-related prefixes were significantly prone to elicit higher biased results than the White groups. Yet, Asian groups did not experience higher levels of algorithmic bias than White people. For sexual orientation, homosexuality-related prefixes were more predictive of highly biased results than heterosexuality in all platforms. Overall, the algorithmic bias of gender, race, and sexual orientation reflected in the search engine's autocomplete predictions remains consistent with the long-standing problems of discrimination and stigma in human society. In fact, it is not just stereotypes and discrimination can be logically and legitimately inherited from society and conveyed to users through algorithms; researchers have found that conspiratorial information has been prevalent in web search results (Urman et al., 2022) and autocomplete predictions (Houli et al., 2021). We call for future research to inspect more diverse kinds of stereotypes or misinformation embedded in algorithms, particularly in the context of autocomplete features.

Our findings also revealed the effects of topic categories vary for different sensitive attributes. For gender, women are more likely to encounter bias in behavioral, appearance & physical, identity, and career topics. For race, the differences between subgroups are more likely to be on topics such as career, appearance & physical. For sexual orientation, identity and cognitive are the main topics that generate algorithmic bias. In addition, we also found autocomplete algorithmic bias related to certain negative attributes, for example, "identity attack" and "insult" were more prominent in SEAPs, which means disadvantaged individuals or groups might suffer from high levels of insulting or inflammatory words, and hateful speech targeting their identity. The analysis in terms of topic categories and various negative attributes goes beyond the simple criticism of inequality and explores the effect of conditional discourse with finer granularity, which provides insights into the representation of cognitive stereotypical bias in algorithms.

The similarities and disparities between platforms provide implications in two aspects. On the one hand, in line with previous literature (e.g., Al-Abbas et al., 2020), we likewise validated the prevalence of search engine autocomplete algorithmic bias in multilingual settings (English and Chinese). On the other hand, disparities between platforms show severer negative algorithmic bias on Baidu than Google and Bing for gender and race. Under the circumstance of media criticism and social monitoring, search engines in English environment (e.g., Google) have responsively optimized the rules related to autocomplete algorithms (e.g., Gibbs, 2016). Baidu, however, does not have those external motivations, as the media and even academia in the Chinese context pay less attention to the autocomplete algorithmic bias. Meanwhile, there is an abnormality against sexual orientation. Wijnhoven and van Haren (2021) concluded that search engine biases may arise and accumulate at multiple points from publishing bias onwards. One possible explanation for the abnormality is the relative scarcity of relevant content in the platform corpus, due to low popularity and visibility of topics related to sexual orientation in the Chinese Internet environment. More detailed investigations could be further explored in future studies.

5.2. Algorithmic bias, autocomplete and search engine mediatization

Ethical concerns associated with using artificial intelligence (AI) are of extremely significance especially considering the "technological unconsciousness" of most users (Beer, 2009: 987-988). Scholars have addressed a wide range of technology applications, even the cutting-edge ChatGPT, to discuss the potential algorithmic bias and reliability issues (Zhuo et al., 2023). The findings of this paper contribute to the research of algorithmic ethics by promoting the awareness and verifying the prevalence of autocomplete algorithmic bias, which is particularly warranted in the context of the widespread of the autocomplete feature in digital applications such as the short video app, TikTok.[2]

Moreover, the specificity of autocomplete feature and its extended forms in emerging digital tools should be highlighted. Firstly, comparing to biases of search engine results page which has gained much attention, affordances of autocomplete feature and relevant users' interaction practices are different. To intervene stereotypes in the search engine results page, individuals or organizations need to create quality-guaranteed contents or employ strategies of search engine optimization (SEO) to get recommendations from ranking algorithms. However, searching by simply entering queries is a behavior with a low threshold. With the ease of use and privacy of search queries, personal biases that would not be expressed under social norms in face-to-face scenarios may be input in the search box and thus, accumulated in the autocomplete algorithm corpora. In addition, the emerging ChatGPT is increasingly functioning as a search engine for information retrieval. Recently, the autocompletion of ChatGPT prompts has been developed (Ezzati, 2023), which risks algorithmic bias since it might complete or reframe our questions based on biased algorithms. More generally, ChatGPT can be recognized as an extended version of search engine autocompletion since it further "completes" answers to users' questions. Unlike search engines' result pages that present query-related items with clear sources for users to refer to, ChatGPT directly generates seemingly credible and knowledgeable responses without citations or references to the original source (Shen et al., 2023). This could render users unconsciously accept the algorithm outputs and take the "standard answer" to a question without thorough examination. Studies have already found that ChatGPT expresses biased opinions about groups and even biased answers to some political-related questions (Motoki et al., 2023; Zhuo et al., 2023). If embraced without thinking, ChatGPT can be prone to biasing users by completing incorrect answers to their questions.

Drawing from the theoretical perspective of mediatization, the autocomplete algorithmic bias implies one distinct aspect of search engine mediatization: the semantic biasing. Andersen (2018) discussed the search engine mediatization in terms of how ideas like ordering become one means of making sense of the mediated construction of reality, with algorithms as one of the influential agents. Algorithms are powerful as they calculate and determine what is hot or trending for users (Gillespie, 2014). What we focus on is the power structures algorithms may generate (Beer, 2009), especially through their semantically biased outputs, such as derogating minorities by tagging them with negative beliefs, or reproducing social inequality by linking women to stereotypical norms. It is indicated that platforms could be the amplifiers and manufacturers of biased discourse through their affordances and users' appropriation (Matamoros-Fernández, 2017). The autocomplete algorithm, by constantly accumulating social stereotypes from users and exposing them to semantically biased predictions, may convey misleading mindsets and amplify their misbeliefs toward specific objects or groups. In this way, semantic biasing becomes a further illustration of search engine mediatization through constructing and mediating our perception of social reality. Future works could further discuss how semantically biased contents affect users in more concrete ways and scenarios.

5.3. Bridging autocomplete algorithmic bias with digital inequality

This paper confirms that the target population of algorithmic bias is consistent with groups in disadvantaged positions, which motivates us to advance the understanding of autocomplete algorithmic bias from the perspective of digital inequality in terms of the underlying mechanism as well as the potential consequences.

Dating back to the production chain, we argue users' "passive participation" in the operation of algorithms acts as one of the mechanisms of autocomplete algorithmic bias. "Passive participation" is described as such where users are not aware of their participation or are drawn to unwilling participation (Lutz and Hoffmann, 2017). Since users' digital footprints are used as training data, particular population groups can suffer from negative algorithm outcomes owing to "passive participation" such as likes, comments, or searches associated with them by others (Lutz and Hoffmann, 2017). Disadvantaged groups or those "at the margins" of society are often the targets of algorithmic bias as data made available about them could more possibly be undesirable and problematic (Marwick et al., 2018; Micheli et al., 2018). In search engine autocomplete, for example, discriminatory queries about females, the Black, and homosexuals are more likely to be typed by some completely different individuals or groups such as racists or homophobia. The relatively vulnerable position of women, homosexuals, and any other minority groups in society renders them easier targets for problematic predictions in autocomplete algorithms.

For the consequences, what is worth noting is the potential negative impacts of discriminatory autocomplete predictions on disadvantaged individuals or groups. More explicitly, the dignity and autonomy of the search target can be compromised by the spread of misinformation (Miller and Record, 2017). That is, autocomplete algorithmic bias may affect the psychological well-being of

[2] A New York Times report observed that "for Generation Z, the video app is increasingly a search engine, too". We tested the search of "why women" in Douyin (the Chinese version of TikTok) in December 2022, and it was also found that the autocomplete results like "(why women) are always angry" and "(why women) cheat more than man". The New York Times report available at: https://www.nytimes.com/2022/09/16/technology/gen-z-tiktok-search-engine.html (accessed February 20, 2023).

vulnerable groups who suffer more severe discrimination through demeaning, stigmatizing, etc.

Moreover, the autocomplete feature might interactively affect users' inquiry by leading to paths they might not have pursued otherwise (Miller and Record, 2017). Although this will not necessarily bring about negative consequences, the practices of using autocomplete algorithms to nudge searches towards unethical directions or results damaging others' reputation still need to be alerted (Lapowsky, 2018), especially at a time when cybercrime and social bot-based political manipulation are prevalent. Since discriminations and stereotypes are once again made visible through presence and nudging of autocomplete predictions, they may contribute to the perpetuation of exhibiting biases in society and the reproduction of social inequalities.

In confronting the negative influences of autocomplete algorithmic bias, more systematic research of online information bias and diverse technological tools are urgently needed. The method of information triangulation, which compares the online information of different data, investigators, theories, and methods (Wijnhoven, 2012), is among existing attempts aiming at evaluating the information trustworthiness and unmasking semantic bias. The anti-thesis searches, as an innovative way for searching queries that have opposing standpoint to detect information bias (Wijnhoven and Brinkhuis, 2015), could also be developed to further examine the biases in queries and autocomplete predictions. Yet future development still requires more debiasing methods and tools. Meanwhile, multiple stakeholders should take their roles and responsibilities in participating in the governance of biased information online. For example, scientists and engineers should endeavor to promote the development of explainable and responsible artificial intelligence. Search engine platforms, apart from actively correcting bias on their own, should also establish clear rules for autocomplete feature and adopt beneficial feedback at the same time. Search engine users could take part in the daily algorithm auditing by improving their awareness of algorithmic bias and reporting biased autocomplete predictions.

Our study contributes to the research of autocomplete algorithm auditing and the measurement of algorithmic bias. In simulating user search query prefixes to elicit stereotyping autocomplete predictions, we advance existing research (Baker and Potts, 2013) by extending the possibility of efficiently collecting SEAPs in declarative sentences with stereotype-sensitive concepts. More importantly, we introduce an innovative approach to measure algorithmic bias in degree based on a set of negative attributes with the help of a machine learning model which measures text toxicity. This contribution will considerably facilitate the identification of more detailed empirical evidence of algorithmic bias from large-scale data. Considering the prevalence of multiple algorithm-driven tools and the escalating influences of ChatGPT, the method proposed in this paper could be further used to test other algorithms, for example, measuring biases in generated answers of ChatGPT using toxicity scores. Various bias detection methods are also needed to deal with emerging problems such as biases in multimodal contents and deepfakes.

6. Limitations

One of the limitations of this paper is that the volume and diversity of retrieved autocomplete results are still limited compared to the giant daily searching logs, thus could not fully reveal the problem of algorithmic bias. Second, the use of Perspective API may not perfectly reflect the algorithmic bias in autocomplete since it is another algorithm by nature. The Perspective API may still maintain some biases through other embedding representations (Gonen and Goldberg, 2019) from data generation to model training. For example, we found that certain SEAPs containing only group identity-related words were given high toxicity scores. Future works could further research the problems with these kinds of debiasing algorithms and develop more advanced approaches for algorithmic bias detection based on cross-validation. Finally, algorithms are continuously evolving and the platforms might modify and revise their autocomplete algorithms at any time. What this paper seeks to do is not merely to verify the existence of autocomplete algorithmic bias, but to highlight more the particularities of this feature, such as its potential unconsciousness by a majority of users, and how its technical affordances can drive hidden stereotypes and discrimination in life to become visible again. Future studies could further explore how autocomplete algorithmic bias impacts users' attitudes and behaviors to unpack substantive effects of autocomplete feature.

Funding information

Chinese National Social Science Major Project: Comprehensive Evaluation and Risk Management of Group Panic Communication under Multiple Emergencies; Project Approval No.: 22&ZD310

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tele.2023.102068.

References

Al-Abbas, L.S., Haider, A.S., Hussein, R.F., 2020. Google autocomplete search algorithms and the Arabs' perspectives on gender: A case study of Google Egypt. GEMA Online® J. Lang. Stud. 20 (4), 95–112. https://doi.org/10.17576/gema-2020-2004-06.

Andersen, J., 2018. Archiving, ordering, and searching: search engines, algorithms, databases, and deep mediatization. Media Cult. Soc. 40 (8), 1135–1150.

Angwin, J., Larson, J., Mattu, S., Kirchner, L., 2016. Machine bias. In: Martin, K. (Ed.), Ethics of Data and Analytics. Auerbach Publications, New York, pp. 254–264.

Azzopardi, L., 2021. Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, pp. 27-37.

Baker, P., Potts, A., 2013. "Why do white people have thin lips?" Google and the perpetuation of stereotypes via auto-complete search forms. Crit. Discourse Stud. 10 (2), 187–204.

Beer, D., 2009. Power through the algorithm? Participatory web cultures and the technological unconscious. New Media Soc. 11 (6), 985–1002.

Behm-Morawitz, E., Mastro, D., 2009. The effects of the sexualization of female video game characters on gender stereotyping and female self-concept. Sex Roles 61, 808–823.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., et al., 2019. Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2212-2220.

Bigman, Y.E., Wilson, D., Arnestad, M.N., Waytz, A., Gray, K., 2022. Algorithmic discrimination causes less moral outrage than human discrimination. J. Exp. Psychol. 152 (1), 4–27.

Bonart, M., Samokhina, A., Heisenberg, G., Schaer, P., 2019. An investigation of biases in web search engine query suggestions. Online Inf. Rev. 44 (2), 365–381.

Büchi, M., Festic, N., Latzer, M., 2018. How social well-being is affected by digital inequalities. Int. J. Commun. 12, 3686–3706.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., He, X., 2020. Bias and debias in recommender system: A survey and future directions. ACM Transactions on Information Systems 41 (3), 1–39.

Couldry, N., Hepp, A., 2017. The Mediated Construction of Reality. Polity Press, Cambridge.

Deho, O.B., Zhan, C., Li, J., Liu, J., Liu, L., Duy Le, T., 2022. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? Br. J. Educ. Technol. 53 (4), 822–843.

Dennhag, I., Steinvall, A., Hakelind, C., Deutschmann, M., 2019. Exploring gender stereotypes about interpersonal behavior and personality factors using digital matched-guise techniques. Soc. Behav. Pers. 47 (8), 1–13.

Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L., 2018. Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 67-73.

Elers, S., 2014. Maori are scum, stupid, lazy: Maori according to Google. Te Kaharoa 7, 1.

Epstein, R., Robertson, R.E., 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proc. Natl. Acad. Sci. 112 (33), E4512–E4521.

Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017, May). "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In: Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 62-71).

Eubanks, V., 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

Ezzati, S., 2023. Superpower ChatGPT 5.0.0 is Released. Retrieved from: https://www.superpowerdaily.com/p/superpower-chatgpt-5-0-0.

Favaretto, M., De Clercq, E., Elger, B.S., 2019. Big Data and discrimination: perils, promises and solutions. A Systematic Review. J. Big Data 6 (1), 1–27.

Friedman, B., Nissenbaum, H., 1996. Bias in computer systems. ACM Transactions on Information Systems (TOIS) 14 (3), 330–347.

Gibbs, S., 2016. Google alters search autocomplete to remove 'are Jews evil' suggestion. Retrieved from: The Guardian. https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocompleteremove-are-jews-evil-suggestion.

Gillespie, T., 2014. The relevance of algorithms. In: Gillespie, T., Boczkowski, P.J., Foot, K.A. (Eds.), Media Technologies: Essays on Communication, Materiality, and Society. MIT Press, Cambridge, MA, pp. 167–193.

Givskov, C., Deuze, M., 2018. Researching new media and social diversity in later life. New Media Soc. 20 (1), 399–412.

Gonen, H., Goldberg, Y., 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862.

Graham, R., 2023. The ethical dimensions of Google autocomplete. Big Data Soc. 10 (1), 20539517231156518.

Hajian, S., Domingo-Ferrer, J., 2012. A methodology for direct and indirect discrimination prevention in data mining. IEEE Transactions on Knowledge and Data Engineering 25 (7), 1445–1459.

Hazen, T. J., Olteanu, A., Kazai, G., Diaz, F., Golebiewski, M., 2020. On the social and technical challenges of web search autosuggestion moderation, arXiv preprint arXiv:2007.05039.

Hepp, A., 2013. Cultures of mediatization. John Wiley & Sons.

Hepp, A., Hasebrink, U., 2014. Human interaction and communicative figurations: The transformation of mediatized cultures and societies. Mediatization of Communication 249–272.

Hosseini, H., Kannan, S., Zhang, B., Poovendran, R., 2017. Deceiving Google's Perspective API built for detecting toxic comments, arXiv preprint arXiv:1702.08138.

Houli, D., Radford, M. L., Singh, V. K., 2021. "COVID19 is_": The Perpetuation of Coronavirus Conspiracy Theories via Google Autocomplete. In: Proceedings of the Association for Information Science and Technology 58 (1), 218-229.

Kasperkevic, J., 2015. Google says sorry for racist auto-tag in photo app. The Guardian. Retrieved from: https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app (accessed 9 April 2023).

Kato, M.P., Sakai, T., Tanaka, K., 2013. When do people use query suggestion? A query suggestion log analysis. Inf. Retr. 16, 725–746.

Kitchin, R., 2019. Thinking critically about and researching algorithms. In: The Social Power of Algorithms. Routledge, pp. 14–29.

Kordzadeh, N., Ghasemaghaei, M., 2022. Algorithmic bias: review, synthesis, and future research directions. Eur. J. Inf. Syst. 31 (3), 388–409.

Krieg, K., Parada-Cabaleiro, E., Medicus, G., Lesota, O., Schedl, M., Rekabsaz, N., 2022. Grepbiasir: A dataset for investigating gender representation-bias in information retrieval results, arXiv preprint arXiv:2201.07754.

Lambrecht, A., Tucker, C., 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manage. Sci. 65 (7), 2966–2981.

Lapowsky, I., 2018. Google autocomplete still makes vile suggestions. Wired. Retrieved from: https://www.wired.com/story/google-autocomplete-vile-suggestions/ (accessed 9 April 2023).

Lundby, K., 2009. Mediatization: Concept, Changes, Consequences. Peter Lang, New York.

Lupton, D., 2015. Quantified sex: a critical analysis of sexual and reproductive self-tracking using apps. Cult. Health Sex. 17 (4), 440–453.

Lutz, C., 2019. Digital inequalities in the age of artificial intelligence and big data. Hum. Behav. Emerg. Technol. 1 (2), 141–148.

Lutz, C., Hoffmann, C.P., 2017. The dark side of online participation: exploring non-, passive and negative participation. Inform. Commun. Soc. 20 (6), 876–897.

Mager, A., 2012. Algorithmic ideology: how capitalist society shapes search engines. Inform. Commun. Soc. 15 (5), 769–787.

Mahdawi, A., 2013. Google's autocomplete spells out our darkest thoughts. The Guardian. Retrieved from: https://www.theguardian.com/commentisfree/2013/oct/22/google-autocomplete-un-women-ad-discrimination-algorithms (accessed 9 April 2023).

Marwick, A.E., et al., 2018. Privacy at the margins— understanding privacy at the margins—introduction. Int. J. Commun. 12, 9.

Matamoros-Fernández, A., 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter. Facebook and YouTube. Inform. Commun. Soc. 20 (6), 930–946.

Mayson, S.G., 2018. Bias in, bias out. Yale Law J. 128 (8), 2218–2300.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54 (6), 1–35.

Micheli, M., Lutz, C., Bûchi, M., 2018. Digital footprints: an emerging dimension of digital inequality. J. Inform. Commun. Ethics Soc. 16 (3), 242–251.
Miller, B., Record, I., 2017. Responsible epistemic technologies: A social-epistemological analysis of autocompleted web search. New Media Soc. 19 (12), 1945–1963.
Motoki, F., Neto, V.P., Rodrigues, V., 2023. More human than human: Measuring ChatGPT political bias. Public Choice 1–21.
Nielsen, R.C., et al., 2017. Social media monitoring of discrimination and HIV testing in Brazil, 2014–2015. AIDS Behav. 21, 114–120.
Noble, S.U., 2018. Algorithms of Oppression. New York University Press.
Olteanu, A., Diaz, F., Kazai, G., 2020. When are search completion suggestions problematic? In: Proceedings of the ACM on Human-Computer Interaction, 4 (CSCW2), pp. 1-25.
Otterbacher, J., Checco, A., Demartini, G., & Clough, P. (2018, June). Investigating user perception of gender bias in image search: the role of sexism. In: The 41st International ACM SIGIR conference on research & development in information retrieval, pp. 933-936.
Rambachan, A., Roth, J., 2020. Design-based uncertainty for quasi-experiments, arXiv preprint arXiv:2008.00602.
Robinson, L., Schulz, J., Dunn, H.S., Casilli, A.A., Tubaro, P., Carvath, R., Chen, W., Wiest, J.B., Dodel, M., Stern, M.J., Ball, C., Huang, K.-T., Blank, G., Ragnedda, M., Ono, H., Hogan, B., Mesch, G.S., Cotten, S.R., Kretchmer, S.B., Hale, T.M., Drabowicz, T., Yan, P., Wellman, B., Harper, M.-G., Quan-Haase, A., Khilnani, A., 2020. Digital inequalities 3.0: emergent inequalities in the information age. First Monday 25, 7.
Roy, S., Ayalon, L., 2020. Age and gender stereotypes reflected in Google's "autocomplete" function: The portrayal and possible spread of societal stereotypes. Gerontologist 60 (6), 1020–1028.
Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C., 2014. Auditing algorithms: Research methods for detecting discrimination on Internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry 22, 4349–4357.
Shen, Y., Heacock, L., Elias, J., Hentel, K.D., Reig, B., Shih, G., Moy, L., 2023. ChatGPT and other large language models are double-edged swords. Radiology 307 (2), e230163.
Shih, M., Pittinsky, T.L., Ambady, N., 1999. Stereotype susceptibility: Identity salience and shifts in quantitative performance. Psychol. Sci. 10 (1), 80–83.
Sullivan, D., 2018. How Google autocomplete works in Search, Retrieved from: https://blog.google/products/search/how-google-autocomplete-works-search/.
Sun, C., Qiu, X., Xu, Y., Huang, X., 2019. How to fine-tune BERT for text classification?. In: In: China National Conference on Chinese Computational Linguistics. Springer, pp. 194–206.
Tahery, S., Farzi, S., 2020. Customized query auto-completion and suggestion—a review. Inform. Syst. 87, 101415.
Urman, A., Makhortykh, M., Ulloa, R., Kulshrestha, J., 2022. Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. Telematics Inform. 72, 101860.
Wang, P., Mi, X., Liao, X., Wang, X., Yuan, K., Qian, F., et al., 2018. Game of missuggestions: Semantic analysis of search-autocomplete manipulations. In Network and Distributed Systems Security (NDSS) Symposium. http://dx.doi.org/10.14722/ndss.2018.23036.
Wijnhoven, F., 2012. The Hegelian inquiring system and a critical triangulation tool for the Internet information slave: A design science study. J. Am. Soc. Inf. Sci. Technol. 63 (6), 1168–1182.
Wijnhoven, F., Brinkhuis, M., 2015. Internet information triangulation: Design theory and prototype evaluation. J. Assoc. Inf. Sci. Technol. 66 (4), 684–701.
Wijnhoven, F., Van Haren, J., 2021. Search engine gender bias. Front. Big Data 4, 29.
Yew, G.C.K., 2019. Search engines and internet defamation: Of publication and legal responsibility. Comput. Law Secur. Rev. 35 (3), 330–343.
Zhang, G., Bai, B., Zhang, J., Bai, K., Zhu, C., Zhao, T., 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5-10 July. 4134–4145.
Zhuo, T.Y., Huang, Y., Chen, C. and Xing, Z., 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. arXiv preprint arXiv:2301.12867.

Cong Lin is a master student at the School of Journalism and Communication at Renmin University of China. He got his bachelor's degree in sociology from Renmin University of China. His academic interests include computational social science, algorithm studies, visual communication, and social inequality.

Yuxin Gao is a master student in the University of Toronto, Canada. Her academic interests include fairness in interaction design, visual communication and human-machine interaction.

Dr. Na Ta is an associate professor at the School of Journalism and Communication at the Renmin University of China. She got her Ph.D. degree in Computer Science and Technology from Tsinghua University. Her research interests include online social networks, platformization and new media, computational communication, and intelligent communication (algorithms, agents, etc.).

Dr. Kaiyu Li is a postdoctoral researcher at the School of Information Technology at the York University. He got his Ph.D. degree in Computer Science and Technology from Tsinghua University. He is interested in the domain of computational communication, data product pricing, data valuation, and data acquisition in data market.

Hongyao Fu a master student at the School of Journalism and Communication at Renmin University of China. Her academic interests include computational social science, human-machine interaction.