Which Stereotypes Are Moderated and Under-Moderated in Search Engine Autocompletion?

Leidinger, A.; Rogers, R.

DOI
10.1145/3593013.3594062

Publication date
2023

Document Version
Final published version

Published in
FAccT '23

License
CC BY

Which Stereotypes Are Moderated and Under-Moderated in Search Engine Autocompletion?

Alina Leidinger
University of Amsterdam
Institute for Logic, Language and Computation
Amsterdam, The Netherlands
a.j.leidinger@uva.nl

Richard Rogers
University of Amsterdam
Department of Media Studies
Amsterdam, The Netherlands
r.a.rogers@uva.nl

ABSTRACT

Warning: *This paper contains content that may be offensive or upsetting.*

Language technologies that perpetuate stereotypes actively cement social hierarchies. This study enquires into the moderation of stereotypes in autocompletion results by Google, DuckDuckGo and Yahoo! We investigate the moderation of derogatory stereotypes for social groups, examining the content and sentiment of the autocompletions. We thereby demonstrate which categories are highly moderated (i.e., sexual orientation, religious affiliation, political groups and communities or peoples) and which less so (age and gender), both overall and per engine. We found that under-moderated categories contain results with negative sentiment and derogatory stereotypes. We also identify distinctive moderation strategies per engine, with Google and DuckDuckGo moderating greatly and Yahoo! being more permissive. The research has implications for both moderation of stereotypes in commercial autocompletion tools, as well as large language models in NLP, particularly the question of the content deserving of moderation.

CCS CONCEPTS

• **Information systems** → **Web search engines**; • **Applied computing** → **Arts and humanities**.

KEYWORDS

stereotypes, search engine autocompletion, Google Autocompletion, content moderation, debiasing, natural language generation

ACM Reference Format:
Alina Leidinger and Richard Rogers. 2023. Which Stereotypes Are Moderated and Under-Moderated in Search Engine Autocompletion?. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3593013.3594062

1 INTRODUCTION

Stereotypes have been defined by Lippmann [40] as 'pictures in our heads' and in the context of research on Google autocompletion as especially reductionist, narrowing 'a person or thing to [specific] traits while exaggerating them' [4]. Is Google, however inadvertently, an engine for perpetuating stereotypes and neglecting moderation? How does it compare to other search engines such as Yahoo! and DuckDuckGo? Stereotypes that are reproduced in language generation or an autocompletion task constitute representational harm [6, 8, 14]. Cadwalladr [10], who made some of the earliest, influential findings concerning Google autocompletion, pointed out that autocompletion stereotypes 'frame' and also 'distort' how we see the world. Noble [48] went further, arguing that stereotypical and racist results perpetuate 'oppressive social relationships'. Indeed, Vlasceanu and Amodio [67] demonstrate that exposure to biased Google image search results reinforces gender stereotyping in professional contexts. Roy et al. [56] link exposure to autocompletions to the psychological process of 'incidental learning' [62] by which information is 'picked up' unintentionally and subconsciously often in the course of another information-seeking activity. Relatedly, Miller and Record [45] argue that autocompletions 'induce changes in epistemic actions', some of which can be harmful [45], especially when stereotypes provide 'ideological

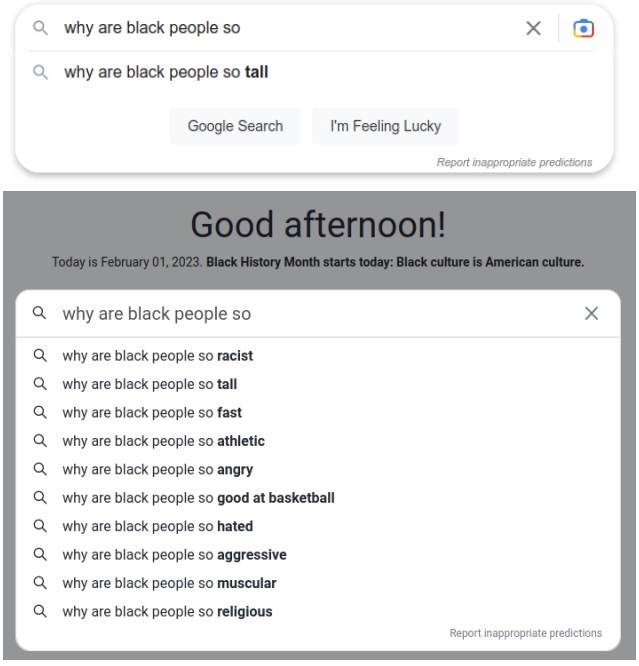

Figure 1: Autocompletions by Google (top) and Yahoo! (bottom) on Feb 1st 2023. Screenshots by authors.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594062

justification' to maintain social hierarchies and further marginalisation [8].

Google is somewhat vague about how autocompletion content moderation works, stating that: "our systems aim to prevent policy-violating predictions from appearing. But if any such predictions do get past our systems, and we're made aware (such as through public reporting options), our enforcement teams work to review and remove them, as appropriate. In these cases, we remove both the specific prediction in question and often use pattern-matching and other methods to catch closely-related variations" [65]. Yahoo! is less expansive than Google in its autocompletion content moderation policy when concerning marginalised groups, removing suggestions when there is hate rather than merely offensive speech [73]. DuckDuckGo does not specify an autocompletion policy, apart from in press reports that the company blocks offensive returns; DuckDuckGo licenses its autosuggestions from Yahoo! [31].

In the following we analyse moderation practices in search query autocompletion, a task common to search engines' proprietary autocompletion algorithms and publicly available, state-of-the-art language models [58, 59, 75]. Given the active, but rather opaque moderation of Google's autocompletion and that of the other engines, in this study we pose the following **research questions**.

(1) For which social groups, are autocompletions suppressed or otherwise moderated in Google, Yahoo! and DuckDuckGo autocompletions? (For a full list of the 150 social groups queried see Table 1.)
(2) When there are autocompletions for the groups under study, how is their sentiment characterised? Are they particularly negative?
(3) How may one portray the moderation of stereotypes in autocompletion by each engine? Are certain engines stricter or more permissive than others?

Contributions: We make 'prompting' or stereotype-eliciting queries concerning approximately 150 terms for social groups (see Table 1) in the three engines based broadly on age, gender, lifestyle, political orientation, peoples, religion and sexual orientation, examining the extent of the suppression or other forms of moderation. We undertook the queries, in order to gain a sense of which autocompletions do not complete or otherwise show signs of moderation. We discuss which stereotypes (and categories of social groups) receive which types of suppression or other moderation, thereby charting the work engines are doing to thwart such outputs. Through scoring the groups by moderation of stereotypes, we also shed light on which group stereotypes are considered rather sensitive by a search engine given their removal or editing. We thereby are able to characterise result moderation overall and per search engine under study.

Our findings could inform work on content moderation policy whether in autocompletion or NLP more generally, particularly by drawing attention to under-moderated categories that have negative suggestions. In light of the harmful impact of stereotype perpetuation, we believe public discourse on moderation priorities as well as transparent documentation on the parts of commercial language technology providers to be crucial.

2 RELATED WORK

2.1 Content moderation of search engine autocompletion

Most studies focus on Google's results moderation, rather than other engines', given its market dominance [4, 33, 55]. Content moderation has been defined by Grimmelmann [30] as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse". Previous work on the moderation of especially Google autocompletion has concerned itself with how it was once prone to outputting derogatory content such as 'are Jews [evil]', where the autocompletion part is in brackets [10]. Indeed, journalists and scholars alike have reported particularly shocking outputs for queries of 'women' [72], 'old men' and 'old women' [55], religions [10], sexual orientation [4], gender identity [2] and others.

Generally speaking, up until 2016, Google product outputs, from web search to autocompletion, were described as 'organic' by the company, or reflections, however unpleasant, of 'what was happening on the web' [10]. After press reports there were noticeable, emergency take-down's and patches [25] in autocompletion. Generally, however, results came with disclaimers (in banner ads) and further explanation (in blog posts or in response to the press) concerning how they 'reflected' what was happening 'across the web' [5, 23].

That state of affairs changed with the introduction of the autocompletion feedback tool in 2017, where users could report on content that they considered 'hateful, racist, offensive, vulgar, sexually explicit, harmful, dangerous, violent, misleading or inaccurate' [27]. In 2018, Danny Sullivan, the company's public search liaison, explained in a long blog post the company's autocompletion removal policies [63], pointing to Google's definition of inappropriate content, particularly derogatory output, relating how the engine removes autocompletions that are 'hateful or prejudicial' with respect to 'race, ethnic origin, religion, disability, age, nationality, veteran status, sexual orientation, gender, gender identity, or any other characteristic that's associated with systemic discrimination or marginalisation'.

Behind the need for the moderation of autocompletion (as well as other suggestions or predictions that appear in other search engine products) are the liabilities that arise from outputting words connected to the incipient search query. Do they defame individuals [12]? Could they induce illegal acts such as downloading of copyrighted material [36]? Do they lead to sources of child pornography or other illicit material [18]? Do they contain hateful language towards groups [20]? What Google defines as 'inappropriate content' to be moderated relates directly to these and other legal liabilities [29]. Group stereotyping, however, is more of a grey area, but would fall under the moderation of what Sullivan describes as 'offensive' content [64]. Design choices with respect to moderation of stereotypes are not detailed by the company and constitute the object of study for this work.

We place our work alongside algorithmic auditing [51, 57], platform observability [53], ethical hacking for vulnerabilities as well as (commercial) content moderation critique [26, 54], though each of these approaches has somewhat different emphases.

Table 1: List of groups targeted by stereotypes (English)

Category	Group	Total
Age	boomers, children, kids, millennials, old men, old people, old women, teenagers, teens	9
Gender/gender identity/ gendered	girls, women, men, females, males, boys, boyfriends, girlfriends, stepmothers, stepfathers, ladies, gentlemen, brothers, sisters, mothers, fathers, grandfathers, grandmothers, wives, husbands, schoolgirls, schoolboys, transgenders	23
Lifestyle/way of living	feminists, frat boys, geeks, goths, hippies, hipsters, nerds, punks, sorority girls, celebrities, criminals, homeless people, poor people, rich people	14
Political/politicised	capitalists, communists, conservatives, immigrants, liberals, populists, socialists, Trump supporters	8
Peoples/ethnicities	Africans, Asians, Asian kids, Asian men, Asian parents, Asian women, African Americans, Black Americans, Blacks, Black fathers, Black kids, Black men, Black parents, Black people, Black women, Europeans, Hispanics, Hispanic men, Hispanic women, Latinas, Latinos, Latin people, Native Americans, Whites, White Americans, White kids, White men, White parents, White people, White women, redheads, gingers, blondes	32
Nationalities	Americans, Afghans, Albanians, Arabs, Australians, Austrians, Bengalis, British people, Chileans, Colombians, Dutch people, Egyptians, Ecuadorians, Ethiopians, Finns, French people, Germans, Ghanaians, Greeks, Indians, Indonesians, Iranians, Iraqis, Irish people, Italians, Koreans, Lebanese people, Mexicans, Moroccans, Nepalis, Nigerians, Norwegians, Pakistanis, Polish people, Romanians, Russians, Scots, Somalis, South Africans, Sudanese people, Swedes, Syrians, Taiwanese people, Turkish people, Ukrainians, Venezuelans, Vietnamese people	47
Religion	Atheists, Buddhists, Catholics, Christians, Hindus, Jews, Mormons, Muslims, Protestants, religious people, Sikhs	11
Sexual orientation	asexual people, bisexual people, gay people, homosexuals, lesbians, pansexual people, queer people	7
Total		151

2.2 Content moderation in Language Models

Search Engine Autocompletion is one real-world application of language modelling (LM) or natural language generation (NLG) which has been demonstrated to suffer from undesirable biases [7, 8, 60, 70].

Methodologically, bias has been quantified using intrinsic measures [9, 11, 66] which operate on word embeddings or extrinsic measures that examine how bias manifests itself in downstream tasks such as sentiment analysis [34, 37] or hate speech detection [19, 52]. For measuring stereotypes in particular, bias benchmarks consisting of contrasting sentence pairs, e.g., StereoSet [46] and CrowS-Pairs [47], have been proposed. In open-ended language generation, prompts are often used to assess to what extent LMs yield undesirable output. Various benchmarks such as BOLD [17], HONEST [49], HolisticBias [61] and RealToxicityPrompts [24] exist for this purpose. Choenni et al. [13] prompt language models to assess to what extent they have learnt stereotypes. In contrast to our work, they use search engine autocompletions as a proxy for stereotypes existing in the real world and compare them to LM output.

Early methods for measuring bias and stereotypes have mainly focused on gender [9, 11]. Recently, the field has turned its attention also towards harms against groups based on, e.g., their disability status [35], gender identity [16], race [22, 43] or religion [1, 42].

Mitigation efforts in NLP include debiasing methods which intervene to produce less biased or stereotyping output, e.g., [9, 21, 28, 69, 74]. LM output can also be flagged as harmful using manual inspection [60], lexicons [49] or another pretrained model [3, 44, 60]. Commercial tools that fall into the latter category mainly focus on hate speech and toxicity, less on stereotypes. Perspective API provides scores for toxicity, insult, profanity, identity attack, threat and sexually explicit [3]. OpenAI reports on the content moderation filter for their language models which scores LM output based on the following criteria: hate, self-harm, sexual content and violence [44, 50].

3 METHODS

3.1 Data collection

We collected autocompletions by prompting three leading search engines, Google, DuckDuckGo and Yahoo!, with the query "why are [group] so ..." for a large number of social groups. For the choice of social groups, we drew on lists of groups from Choenni et al. [13] and StereoSet [46], a benchmark commonly used for measuring stereotypes in LMs. It features stereotypes pertaining to 321 target

terms falling into the categories gender, profession, race and religion. (Categories were originally sourced from Wikidata relation triplets [68].) We follow Choenni et al. [13] in extending this list of social groups, but excluded colloquialisms and slurs. We further reorganised the categorisation using Google's list of groups of potentially marginalised (mentioned in the introduction), resulting in the categories age, gender/gender identity/gendered, lifestyle/ways of living, nationalities, peoples/ethnicities, political/politicised, religion and sexual orientation. While most of our categories match up with Google's, the groups listed under lifestyle/ways of living as well as political/politicised fit Google's catch-all category of 'any other characteristic associated with discrimination or marginalisation'. We removed social groups belonging to the professions category, since the great majority of those are not commonly considered as marginalised[1]. See Table 1 for the full list of social groups.

We followed Choenni et al. [13]'s approach in querying the engines' autocompletion services in January and again in August 2022 using the Python library `requests` and thus simulated an anonymous user querying autocompletions[2]. Hence, autosuggestions were not influenced by, e.g., personal search history. Language and country parameters were set to English and the U.S. region, and the browser setting to Chrome. Since autocompletions can also deviate from the exact wording of the prompt we discarded those autocompletions that did not conform to the phrasing "why are [group] so..."[3].

Whereas Baker and Potts [4] employed prompts as 'why do [group]', 'how do [group]', 'what do [group]' and 'where do [group]', finding them fruitful in triggering stereotypes, our prompt directly elicits stereotypes, as it asks for the reason behind a group's characteristics, thereby assuming those inquiring are not questioning the stereotype, or if questioning it (through sarcasm) are familiar with the stereotype. We release our data as part of the supplementary material of this work [39].

3.2 Analysis of Moderation Practices

To uncover which search engine moderates which target category, we considered the following as strong indicators: 1) The target category contains a large percentage of groups yielding 0 autosuggestions. 2) On average, the number of autosuggestions from this engine is substantially lower than it is for the other search engines (for this category). 3) Common (negative) stereotypes are absent among autosuggestions that do appear in the autosuggestions from other engines. Additionally, we have observed on occasion a number of autocompletions, or single autocompletion, charged with positive sentiment, in comparison to other engines that return many mainly negative autosuggestions (for this category).

We recorded summary statistics and sentiment scores to operationalise our reasoning. To corroborate our findings, we drew a comparison of summary statistics and scores between our two timestamps of January and August 2022. To quantify sentiment, we scored the sentiment of each full autocompletion using a large language model fine-tuned for sentiment classification. Specifically, we used RoBERTa [41] optimised by Hartmann et al. [32] for this purpose[4]. We chose this model in particular, since it is fine-tuned on a large set of English language datasets stemming from various domains, e.g., tweets, reviews, etc. Binary sentiment scores are in the range between 0 and 1 with higher scores indicating more negative sentiment.

4 FINDINGS

In the following we discuss our findings with respect to the moderation (or under-moderation) of autocompletion by the search engines. Which categories and terms for groups appear to be the source of moderation? When one engine returns plentiful results including (negative) stereotypes, while another returns none or a single result, our findings of moderation are supported. While there are exceptions, we are able to characterise the individual engines, generally, as greatly moderating (Google, DuckDuckGo) and permissive (Yahoo!).

4.1 The moderation of autocompletion

None of the engines returned autocompletions for sexual orientation as a whole. Across the engines the categories nationalities, peoples/ethnicities, political/politicised and religion had relatively few autocompletions. For Google, age and gender autocompletions are overall the most negative, though there are outliers (where there are very few returns and those returns are negative as is the case with 'why are Protestants so' in Google). Yahoo! has overall the most negative autosuggestions, while DuckDuckGo has the least. The overall average sentiment scores for autocompletions are 0.78 for Yahoo!, 0.59 for Google and 0.49 for DuckDuckGo, where the higher the score the more negative the sentiment (for more details see Table 9).

4.1.1 Sexual orientation. None of the search engines, in either time period, served autocompletions for groups in the sexual orientation category. Sexual orientation is alone in this regard, indicating a particularly well moderated set of terms.

4.1.2 Religion. Autcompletions for social groups in the religion category seem to be heavily moderated by Google and DuckDuckGo (See Table 3). Yahoo!, contrariwise, furnishes a substantial number of autocompletions in particular for Jews, including anti-Semitic slurs such as 'cheap' and 'rich'. Google returns no autocompletions for religious groups with the exception of Mormons where we see, in both periods, the potentially actively curated, single suggestion of 'nice'. The other religious groups that are under-moderated, at least for one time period, are Protestants ('bitter', 'boring', 'so-called', 'judgemental') as well as Christians ('judgemental'), which are mainly negative qualifiers and result in a negative sentiment score. DuckDuckGo appears to block all autocompletions for religions. There are no autocompletions for Hindus and Buddhists from any of the search engines.

[1]We would like to add that some intersectional groups, e.g., 'old women', fall into more than one category, i.e., gender and age. We decided to follow the categorisation of Choenni et al. [13] in this case. We could have also created more intersectional categories (e.g., queer Indian men), but left the broader terms (with up to one qualifier) so that it would allow for comparison of moderation attention (across engines) in the categories demarcated by Google.

[2]The source code developed by [13] is available here: https://github.com/RochelleChoenni/stereotypes_in_lms

[3]We found this prompt to be particularly effective in returning the maximal number of results for non-marginalised or non-politicised groups during an initial data exploration, compared to four others in the original research by Choenni et al. [13].

[4]We used the HuggingFace library [71] and the following checkpoint https://huggingface.co/siebert/sentiment-roberta-large-english.

Table 2: Proportion of queries that yield 0 or 1 autosuggestions. US January 2022 (left) August 2022 (right)

Category	# groups	Google	Yahoo!	Duck.	Google	Yahoo!	Duck.
Age	9	50%	25%	25%	50%	25%	25%
Gender/gender identity/gendered	23	39.1%	43.5%	60.9%	30.4%	47.8%	60.9%
Lifestyle/way of living	14	50%	78.6%	78.6%	57.1%	78.6%	78.6%
Political/politicised	8	100%	62.5%	87.5%	100%	62.5%	87.5%
Peoples/ethnicities	32	75.8%	42.4%	87.9%	84.8%	48.4%	87.9%
Nationalities	47	78.8%	74.5%	66%	78.8%	85.1%	66%
Religion	11	100%	36.4%	100%	90.9%	36.4%	100%
Sexual orientation	7	100%	100%	100%	100%	100%	100%

Table 3: Autocompletions for religious groups, US Jan. and Aug. 2022, where autocompletions in normal font are from both Jan. and Aug., bold autocompletions are from Jan. only and italicised from Aug. only

Group	Google	Yahoo!	D.
Atheists	-	afraid of God, angry	-
Catholics	-	against abortion, liberal, *negative, unlike christ, into politics,* **devoted to Mary**	-
Christians	*judgemental*	angry, controlling, divided, easily offended, fearful, happy, hated, *judgemental to gays*	-
Jews	-	liberal, persecuted, powerful, rich, cheap, smart, successful, hated, wealthy, *funny*, **disliked**	-
Mormons	nice	happy, interested in genealogy, successful, prepared, rich, strict, *wealthy, patriotic, controversial*, **interested in ancestry, into genealogy, misunderstood**	-
Muslims	-	religious, spoken word, *conservative*	-
Protestants	*bitter, boring, so-called, judgemental*	divided	-
Religious people	-	racist, brainwashed, *miserable*, **negative**	-

Table 4: Autocompletions for political/politicised groups, US, Jan. and Aug. 2022, where autocompletions in normal font are from both Jan. and Aug., bold autocompletions are from Jan. only and italicised from Aug. only

Group	Google	Yahoo!	Duck.
Immigrants	successful	-	-
Trump supporters	-	angry, **brainwashed**, delusional, gullible, hateful, ignorant, loyal, **mad**, stupid, violent, *dumb fat*	-
Conservatives	afraid of higher education	hateful, angry, miserable, racist, brainwashed, stubborn, intolerant, anti abortion, *paranoid, mean*, **cold hearted, fearful**	afraid of change, *pro life*
Liberals	popular in Canada	angry, condescending, **dumb**, hateful, ignorant, intolerant, racist, stupid, unhappy, violent, *miserable*	-

4.1.3 Political/Politicised. DuckDuckGo and Google seem to be moderating autocompletions for groups in this category quite rigorously, while for Yahoo! the content management is less prevalent (Table 4). DuckDuckGo blocks nearly all autocompletions. For Google autocompletions are mainly suppressed, with one exception, conservatives, who are 'afraid of higher education'. In Yahoo! conservatives are 'hateful', 'angry', 'miserable', 'brainwashed', etc. as are liberals, resulting in a high negative sentiment score. No search engine provided any autocompletions for communists, socialists, capitalists or populists.

Table 5: Autocompletions for select peoples/ethnicities, US, Jan. and Aug. 2022, where autocompletions in normal font are from both Jan. and Aug., bold autocompletions are from Jan. only and italicised from Aug. only

Group	Google	Yahoo!	D.
Black people	good at running, *tall*	aggressive, angry, athletic, **big**, cool, fast, hated, racist, religious, tall, *good at sports*	-
Black Americans	-	much bigger than Africans today, poor, racist, violent	-
African Americans	-	loud, religious, good at sports, angry, *sensitive about everything*, **big, tall**	-
Blacks	-	racist against Whites, *religious, hated in america, different physically, tall*	-
Black men	-	aggressive, tall, cheap, angry, attracted to white women, violent, *homophobic, popular, hot, promiscuous,* **hated**	-
Black women	-	angry, big, difficult, loud, religious, undesirable, violent, *heavy, hard to date, confident*	-
Black kids	-	*angry.* bad, disrespectful, disrespectful in school, **hyper, violent,** wild	-
Black parents	-	mean, strict, *abusive*	-
White people	-	angry, attractive, bad, beautiful, entitled, racist, rich, scared of Black people, smart, violent	-
White Americans	-	cold, racist	-
Whites	-	angry, afraid of Blacks, obsessed with Blacks, racist, racist against Blacks, smart	-
White men	-	angry, attracted to Asian women, entitled, nice, tall, threatened by Black men, threatened by Black women, violent	-
White women	-	angry, attracted to Black men, beautiful, easy, hot, sensitive, thin	-
White kids	-	bad, smart	-
White parents	laid back	-	-
Asians	smart, good at math	*peaceful*	-
Asian men	-	feminine	-
Asian women	-	popular	-
Asian kids	-	depressed, smart, *short*	-
Asian parents	strict on grades, critical, *negative, judgemental, angry*, **critical reddit, loud**	strict about grades	-
Latin people	-	**stupid**	-
Latinos	-	**involved in Kobe Bryant**, *passionate*	-
Latinas	**hard to date**, *loyal, desired*	*easy*	-
Hispanics	-	loud	-

4.1.4 Peoples & ethnicities. Overall, moderation practices in this category fit the overall pattern of Google and DuckDuckGo as moderating engines and Yahoo! as more permissive (see table 5). Yahoo! returned the largest number of autocompletions which also resulted in the highest score for negative sentiment. Both Google as well as DuckDuckGo returned a lower sentiment score. In particular, for the terms, Black people and White people, we found few autocompletions in Google, a complete lack of DuckDuckGo autocompletions and some strong stereotyping in Yahoo! autocompletions, which accounts for the negative sentiment. Yahoo! autocompletions touch on racism ('why are White people racist') and negative images ('aggressive', 'angry', 'bad', 'mean', etc.). A number of Yahoo! autocompletions we found for Latinos ('loud', 'stupid', 'involved in Kobe Bryant') and Asians (feminine men, popular women, smart and depressed kids, strict parents) contain negative valences.

4.1.5 Nationalities. Google's moderation extends to nationalities. Theirs and DuckDuckGo's autocompletions were among the most positive of all autocompletions we collected. Google outputs positive autocompletions for many nationalities, including Americans ('friendly'), Germans ('smart', 'tall'), Indians ('smart'), Moroccans ('strong', 'beautiful'), Australians ('tall'), Russians ('tall', 'good at chess', 'pretty'), Somalis ('rich', 'tall', 'successful') and Syrians ('beautiful'). It returns no autocompletions for most groups, including Austrians, British people, Ethiopians, French people, Greeks, Irish people, Italians, Mexicans, Nigerians, Pakistanis, Polish people and Romanians. When DuckDuckGo returns autocompletions for

nationalities, there is a smattering of stereotypes in evidence, but not enough to result in a negative sentiment score. Yahoo! again scored highest for negative sentiment. For example, Egyptians are 'loud', French people are 'mean' and Germans are 'cold'.

4.1.6 Gender/gender identity/gendered. None of the search engines returned results for 'transgenders', but otherwise the results in this category are perhaps the most surprising overall. For the remaining social groups we see a substantial number of autocompletions, many of which are stereotypical as well as insulting (see table 7). All three engines scored on the negative end of the sentiment spectrum with Yahoo! being overall the most negative, followed by Google and then DuckDuckGo. Google gives unflattering suggestions for most of the terms in the gender category. For women, females and girls, we found such autocompletions as 'controlling', 'clingy' and 'dramatic', and for men, males and boys 'boring', 'mean', 'insensitive', 'lonely' and 'immature'. Yahoo!'s negative autocompletions are more plentiful. For men, males and boys we found such suggestions as 'difficult', 'complicated', 'angry at women', 'needy', 'insensitive' and 'confusing'. DuckDuckGo furnishes fewer autocompletions than Google and Yahoo!, but the autocompletions have similar terms.

4.1.7 Age. Google is the only engine that curates most autocompletions for the age category as the overall number of autocompletions is comparatively small (see Table 8). It stands alone in suppressing most ageist autocompletions for queries concerning older people, with the exception of old people as 'entitled'. Yahoo! and DuckDuckGo return stereotypes as 'angry', 'grumpy', 'cold', 'negative', 'stubborn', 'difficult', 'entitled' and 'slow'. While Google's is marginally lower, sentiment scores for all three engines combined were among the highest overall. All engines return stereotyping autocompletions for boomers, children and teenagers.

4.1.8 Lifestyle/Ways of living. This category follows the overall pattern of Yahoo! furnishing a large amount of negative autocompletions, Google some (though not for the same ones) and DuckDuckGo returning few. Autocompletions for feminists, the homeless, the rich, the poor and criminals are suppressed by nearly all search engines, and rather exceptionally there is possibly evidence of (positive) moderation on the part DuckDuckGo as well as Yahoo! (see table 6). Yahoo! autocompletions for rich people are in evidence, while poor people (happy, poor) and homeless people (happy) have positive inflections. Google is the only search engine to return completions for punks, frats, goths, hippies, hipsters and nerds. DuckDuckGo and Yahoo! have no results, with the exception of DuckDuckGo's 'why are nerds so [attractive, successful]', which could be an Easter egg.

4.1.9 Autocompletion engine moderation and sentiment. All in all, Google appears to moderate results in much greater quantities than DuckDuckGo and Yahoo!. Google often returns no autocompletions in both January and August 2022 (e.g., for not only sexual orientation, but also others as seen in Table 2) or single results, some charged with positivity, e.g., 'why are immigrants so successful' (see Table 9 for an overview). Given the universe of stereotypes potentially associated with the groups, when only one autocompletion appears (and has a positive valence) it could be an indication of curation, a point we return to in the discussion.

As a rule, through such moderation, the sentiment scores become less negative, compared to Yahoo!. DuckDuckGo mainly suppresses (potentially) stereotypical or inappropriate autocompletions completely and overall has the lowest negative sentiment scores. Yahoo!, characterised as by far the most permissive engine, was found to moderate the least and have the highest negative sentiment. In those cases when it does not permit stereotypes to appear, it removes all autosuggestions.

5 DISCUSSION

We would like to discuss four implications of the research. The first concerns the distribution of moderation overall, the second the permissiveness of particular search engines for certain queries, the third the continuing stakes of perpetuating certain negativity or insults in services that the user cannot turn off, and finally the question of the transparency of the moderation. We also would like to ask whether engines can do better.

In the research we found a hierarchy of concern, which we suggest could be flattened further. We also found a differentiation in moderation across engines, which could be evened. With respect to the hierarchy of concern, sexual orientation is moderated as are most ethnicities and religions, though with some exceptions (such as Protestants in Google). Gender is under-moderated, given the stereotypes and insults returned for especially women. Older people as a category is also under-moderated, at least compared to the other categories. With respect to individual engines, Yahoo!'s moderation stands out for the amount of stereotypes and insults allowed to pass through across most categories. Given that there is the exception of sexual orientation, Yahoo! is not an un-moderated engine but certainly one where attention is called for.

The under-moderation has resulted in negative autocompletions, as evidenced by the sentiment scoring. These are groups of people who historically have faced discrimination and marginalisation, and the autocompletions could be considered what Noble [48] called 'reinforcement'. Users thereby can come across the stereotypes and insults, 'picking them up' while searching for other information, learning abusive remarks for groups or witnessing their reinforcement. Is the presence of these stereotypes and insults reason enough to make the service optional or disabled by default?

When certain groups see stereotypes and insults and others are conspicuously absent, the question arises about search engine policy and its implementation. While there appears to have been an expansion in moderation activities over the past few years, its documentation has been supplied only in rather general terms. While we have read company blog posts concerning the moderation of this content, as far as we can tell the scope as well as the types of moderated stereotypes have not as yet become part of transparency reports or other official company documentation. Moreover, harmful stereotypes are also not among the kinds of inappropriate autocompletion content that users can report through the interface tool of search engines, at least explicitly. Documentation on content filters built into commercial Language Models often does not mention stereotypes explicitly either [50]. The implication is that search engines could provide not only a content moderation policy but also evidence (beyond the blog posts) of its effective implementation.

Table 6: Autocompletions for select lifestyle/way of living groups, US, Jan. and Aug. 2022, where autocompletions in normal font are from both Jan. and Aug., bold autocompletions are from Jan. only and italicised from Aug. only

Group	Google	Yahoo!	DuckDuckGo
feminists	-	angry all the time	-
homeless people	-	**happy**	-
poor people	-	angry, *mad*, **happy loud**	-
rich people	-	cheap, mean, stingy, miserable, liberal, wasteful, entitled	**healthy**, rich

Table 7: Autocompletions for select gender/gender identity/gendered, US, Jan. and Aug. 2022, where autocompletions in normal font are from both Jan. and Aug., bold autocompletions are from Jan. only and italicised from Aug. only

Group	Google	Yahoo!	DuckDuckGo
women	attractive, **beautiful**, controlling, hot, sensitive, **short**	difficult, unhappy, mean to each other, emotional angry, *jealous of other women, important, dramatic, picky about men*	-
females	hot, clingy, emotional during period, *competitive, messy, stubborn, cute, defensive*, **attractive, sensitive, bippolar, weak in naruto**	difficult, emotional, **jealous of me**, protective against other female	moody, emotional during period, attractive, sensitive, emotional, entitled, *cute*, **aggressive**
girls	dramatic, emotional, insecure, sensitive, attractive, *soft*, **pretty, cute**	complicated, dramatic, emotional, difficult, sensitive, short, mean to each other, confusing, *expensive, competitive*, **mean to boys**	cute, confusing, pretty, sensitive, attractive, wierd, mean, *dramatic*, **hot**
men	complicated, boring, sensitive, hot, insensitive, *lonely*, **warm, hairy**	insensitive, visual angry, needy, *attracted to women, loud, hot and cold, complicated, angry at women, sensitive*, **difficult**	shallow, aggressive, self-absorbed, *simple, jealous sensitive, moody*, **hot, cute**
males	attractive, attracted to females, *angry*, **rare, mean, loud**	difficult, emotional, protective against other female	*attractive*, **aggressive**
boys	**cute**, immature	complicated, confusing, mean to girls, stupid quotes, *wearing nail polish, dramatic when sick, loud, competitive, tall*, **cute best friends, sensitive, destructive**	ugh, confusing, aggressive, cute, *funny complicated*, **difficult, hot, strong**

These four points aim to orient the discussion around autocompletion moderation, particularly the decisions on what to moderate as well as disclose about moderation and the stakes of undermoderating.

5.1 Additional categories of moderation strategy?

A cursory look at the press reports in the 2010s concerning particularly shocking autocompletions such as 'are Jews [evil]' yields follow-up articles detailing how those suggestions have been 'fixed' or 'removed' [25]. Such suppressions or patches are implemented in direct response to journalistic discoveries, and related queries are presumably also fixed such as 'are Christians [evil]'.

But we also have observed autocompletion result lists that leave a single suggestion, occasionally charged with positivity, such as 'why are homeless people so happy'. Apart from the blocking, this example could point to a third way which could be dubbed curation, which entails retaining fewer (sometimes positively charged) results.

Curation is more complex, however, when considering autocompletion results that contain synthetic content (such as the query completed with 'near me' or 'meaning') or knowledge base content (such as the query completed names of pop songs, famous people or official organisations). For an overview of examples see table 12 in the appendix.

Prior to Sullivan's blog posts at Google there was not much written about specific moderation practices, especially the interplay in autocompletion between organic results (the output of 'real searches') and synthetic ones (the output of 'word patterns'). In our dataset there are synthetic additions (see Table 12) that did not result in more substantive or useful autocompletions, raising the question of the current effectiveness of the strategy of using such 'word patterns' as a part of content moderation compared to pruning autocompletion outputs.

As Danny Sullivan indicated in the 2020 post, certain pruning, however, could be construed as a form of editing that overly minimises offence. In turn it could result in a politicised, public outcry.

6 CONCLUSION

Overall, autocompletion is an actively moderated space. We found a distribution of moderation of categories, with some intuitive as well as counter-intuitive results. Sexual orientation has been moderated

Table 8: Autocompletions for age, US, Jan. and Aug. 2022, where autocompletions in normal font are from both Jan. and Aug., bold autocompletions are from Jan. only and italicised from Aug. only

Group	Google	Yahoo!	DuckDuckGo
boomers	aggressive, controlling, rich out of touch, *entitled, bad with technology, entitled reddit, out of touch reddit*, **loud, bad with technology reddit, angry reddit, clueless**	toxic, conservative, angry, *annoying*	selfish, fat, conservative, greedy, entitled, *liberal*, **salty**
children	loud, **annoying**	energetic, important, disrespectful, cruel, honest, expensive, annoying, impressionable, *special, competitive*, **stubborn, easily influenced**	curious, creative, loud, important, resilient, noisy, *loving, honest*, **vulnerable, cute**
kids	loud, energetic, cute, happy, *cruel*, **annoying**	annoying, cruel, loud, selfish, stupid, mean, disrespectful, *sensitive these days, dumb today, fat*, **happy, lazy these days, toxic**	weird, energetic, fat, noisy, loud, *cringe, entitled now, mean in middle school* , **cute, mean to other kids, happy**
millennials	-	-	-
old people	*entitled*	cold, tired, angry, negative, *dependent, naive*, **loud**	difficult, grumpy, entitled, stubborn, *grouchy, negative, nice, stiff when they get up*, **cold, slow, cute**
old men	-	-	-
old women	-	-	-
teenagers	angry, *sad*	depressed, difficult, mean to their parents, angry, emotional, hormonal, tired all the time, irritable, *alienating toward each other, forgetful*, **stressed, unhappy**	angry, moody, lazy, tired, emotional, *skinny, awkward*, **stressed, dramatic, grumpy**
teens	stressed, depressed	difficult, attached to their phones, tired, addicted to their phones, *depressed these days, sad derek thompson, sad today, easily influenced*	moody, depressed, emotional, sad, rebellious, stressed, edgy, *impulsive*, **lazy**

out of autocompletion. Peoples and ethnicity are highly moderated, followed by religion. The category gender/gender identity/gendered is rather under-moderated, however, and is populated with stereotypes with negative attributions. Stereotypes are attached to both men and women. Age is also rather under-moderated and autosuggestions are more negative, though Google moderates more than the other engines. There are sporadic stereotypes in all engines, even those as Google (the example of Protestants) and DuckDuckGo (nationalities). One rather counter-intuitive finding is the lack of moderation in Yahoo!. While sexual orientation and a few other sensitive categories have been addressed, compared to Google and DuckDuckGo, the situation with Yahoo! is not that far removed from what Baker and Potts [4] described for Google in 2013: 'The auto-completion questions offer a window into the collective Internet consciousness, and what this window reveals is not an attractive scene.' Indeed, the sentiment associated with Yahoo!'s autocompletions under study is considerably more negative compared to Google's and DuckDuckGo's.

Our work more generally has pointed to moderation (in Google in particular) across a range of terms that were not moderated a decade ago, according to the journalistic pieces where offensive results were reported. Age is under-moderated, with the exception of Google. Overall gender, however, remains under-moderated.

7 LIMITATIONS AND FUTURE WORK

There are several limitations to be discussed, including the work's U.S.-centric orientation, our search engine choice, the formulation of prompts and the use of pretrained language models for sentiment classification. The research undertaken is largely U.S.-centric and certain of the stereotypical sensitivities could be interpreted as such. Future work would benefit from developing culturally specific sets of queries across a variety of languages in order to study moderation practices across regions (and compare regions). Given the U.S.-centric orientation, we also could have included Bing, Ecosia and other smaller engines. Studying Baidu, Yandex and Naver could broaden the scope of comparison.

We re-used the prompts from previous work, remapping them onto Google's categories of derogatory remarks, but could have added ones from other journalistic or scholarly work on autocompletion. While our lists of social groups does cover some intersections [15], they do so with only one qualifier. Finally, there were certain groups for which no results were returned, which one could interpret as moderation or as the result of a relevance threshold

Table 9: Sentiment score (higher score is more negative) US Jan 2022 (left) and Aug 2022 (right) (N: number of completions per category, μ: average sentiment score, σ: standard deviation of sentiment scores)

Category		Google	Yahoo!	Duck.	Google	Yahoo!	Duck.
Age	N	18	42	45	19	48	45
	μ	70.0	83.2	75.3	76.6	85.2	75.5
	σ	0.4	0.4	0.4	0.4	0.4	0.4
Gender/gender identity/gendered	N	98	64	52	90	79	58
	μ	68.3	77.8	59.6	72.9	76.7	65.5
	σ	0.4	0.4	0.5	0.4	0.4	0.5
Lifestyle/way of living	N	31	20	12	30	20	11
	μ	74.9	81.0	43.9	63.4	86	47.7
	σ	0.4	0.4	0.5	0.5	0.3	0.5
Political/politicised	N	3	30	1	3	30	2
	μ	33.5	96.5	99.8	33.5	96.5	99.1
	σ	0.6	0.2	-	0.6	0.2	0.01
Peoples/ethnicities	N	30	94	14	33	106	16
	μ	63.1	73.3	35.8	54.4	71.5	43.6
	σ	0.5	0.4	0.5	0.5	0.4	0.5
Nationalities	N	60	36	88	54	51	89
	μ	28.1	68.4	31.9	32.9	66.2	29.3
	σ	0.4	0.4	0.5	0.5	0.5	0.4
Religion	N	5	38	0	2	42	0
	μ	96.5	75.5	-	91.1	73.2	-
	σ	0.08	0.4	-	0.1	0.4	-
Sexual orientation		-	-	-	-	-	-

of the candidate autocompletion. When one peruses the groups to which this observation applies, the likelihood that they have been moderated (rather than underpopulated with associations) remains high.

Pretrained language models have been shown to suffer from what is termed 'lexical bias', meaning that they associate the mere mention of a marginalised identity with negative sentiment Kiritchenko and Mohammad [38]. This might drive up scores for negative sentiment for certain categories.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their insightful comments. The work for this publication is financially supported by the project, 'From Learning to Meaning: A new approach to Generic Sentences and Implicit Biases' (project number 406.18.TW.007) of the research programme SGW Open Competition, which is (partly) financed by the Dutch Research Council (NWO).

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence* 3, 6 (2021), 461–463.
[2] LS Al-Abbas, Ahmad S Haider, and Riyad F Hussein. 2020. Google autocomplete search algorithms and the Arabs' perspectives on gender: A case study of Google Egypt. *GEMA Online® Journal of Language Studies* 20, 4 (2020), 95–112.
[3] Perspective API. 2023. *About the API - Attributes and Languages.* https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US
[4] 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 (may 2013), 187–204. https://doi.org/10.1080/17405904.2012.744320
[5] Judit Bar-Ilan. 2006. Web links and search engine ranking: The case of Google and the query "jew". *Journal of the American Society for Information Science and Technology* 57, 12 (2006), 1581–1589.
[6] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society.*
[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 610–623.
[8] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 5454–5476.
[9] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in neural information processing systems* 29 (2016), 4349–4357. http://papers.nips.cc/book/advances-in-neural-information-processing-systems-29-2016
[10] Carole Cadwalladr. 2016. Google, democracy and the truth about internet search. *The Guardian* 4, 12 (2016), 2016.
[11] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230
[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, University of Hong Kong Faculty of Law Research Paper* 2015/018 (2015).
[13] Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* 1477–1491.
[14] Kate Crawford. 2017. The trouble with bias. Keynote at NeurIPS.
[15] Kimberlé W Crenshaw. 2017. *On intersectionality: Essential writings.* The New Press.
[16] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges

in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1968–1994.
[17] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 862–872.
[18] Nicholas Diakopoulos. 2015. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism* 3, 3 (2015), 398–415.
[19] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA, 2018-12-27) *(AIES '18)*. Association for Computing Machinery, 67–73. https://doi.org/10.1145/3278721.3278729
[20] Steve Elers. 2014. Maori are scum, stupid, lazy: maori according to Google. *Te Kaharoa* 7, 1 (2014).
[21] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, 2019). Association for Computational Linguistics, 1696–1705. https://doi.org/10.18653/v1/P19-1166
[22] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1905–1925.
[23] LJ Flynn. 2004. Google says it doesn't plan to change search results. *The New York Times* (2004).
[24] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3356–3369.
[25] Samuel Gibbs. 2016. Google alters search autocomplete to remove'are Jews evil'suggestion. *The Guardian* 5 (2016).
[26] Tarleton Gillespie. 2018. *Custodians of the Internet.* Yale University Press.
[27] Ben Gomes. 2017. Our latest quality improvements for Search. *Google Blog* (2017).
[28] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, 2019-06). Association for Computational Linguistics, 609–614. https://doi.org/10.18653/v1/N19-1061
[29] Google. 2022. *Removing Content From Google.* https://support.google.com/legal/troubleshooter/1114905#ts=9814647%2C9815053%2C3337372
[30] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
[31] Kirsten Grind, Sam Schechner, Robert McMillan, and John West. 2019. How Google interferes with its search algorithms and changes your results. *The Wall Street Journal* 15 (2019).
[32] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2022. More than a feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing* (2022).
[33] Timothy J Hazen, Alexandra Olteanu, Gabriella Kazai, Fernando Diaz, and Michael Golebiewski. 2022. On the social and technical challenges of Web search autosuggestion moderation. *First Monday* (2022).
[34] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (Online, 2020-11). Association for Computational Linguistics, 65–83. https://doi.org/10.18653/v1/2020.findings-emnlp.7
[35] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5491–5501.
[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] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (New Orleans, Louisiana, 2018). Association for Computational Linguistics, 43–53. https://doi.org/10.18653/v1/S18-2005
[38] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).
[39] Alina Leidinger and Richard Rogers. 2023. *Stereotype elicitation in Google, DuckDuckGo and Yahoo! autcompletion.* https://doi.org/10.5281/zenodo.7906930
[40] Walter Lippmann. 1922. Public Opinion, New York, McMillan.
[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[42] Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially Aware Bias Measurements for Hindi Language Representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1041–1052.
[43] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 615–621.
[44] Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2022. A holistic approach to undesired content detection in the real world. *arXiv preprint arXiv:2208.03274* (2022).
[45] 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.
[46] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5356–5371.
[47] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1953–1967.
[48] Safiya Umoja Noble. 2018. *Algorithms of oppression.* New York University Press.
[49] Debora Nozza, Federico Bianchi, Dirk Hovy, et al. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
[50] OpenAI. 2023. *Moderation.* https://platform.openai.com/docs/guides/moderation/overview
[51] Devah Pager. 2007. The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *The Annals of the American Academy of Political and Social Science* 609, 1 (2007), 104–133.
[52] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, 2018-10). Association for Computational Linguistics, 2799–2804. https://doi.org/10.18653/v1/D18-1302
[53] Bernhard Rieder and Jeanette Hofmann. 2020. Towards platform observability. *Internet Policy Review* 9, 4 (2020), 1–28.
[54] Sarah T Roberts. 2019. *Behind the screen.* Yale University Press.
[55] Senjooti Roy and Liat Ayalon. 2020. Age and gender stereotypes reflected in Google's "autocomplete" function: The portrayal and possible spread of societal stereotypes. *The Gerontologist* 60, 6 (2020), 1020–1028.
[56] Senjooti Roy, Liat Ayalon, Gabi Weisfeld, and Barbara J Bowers. 2020. Age and Gender Stereotypes Reflected in Google's "Autocomplete" Function: The Portrayal and Possible Spread of Societal Stereotypes. *The Gerontologist* 60, 6 (aug 2020), 1020–1028. https://doi.org/10.1093/GERONT/GNZ172
[57] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014), 4349–4357.
[58] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Scao, Arun Raja, et al. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.
[59] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).
[60] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4275–4293.
[61] Eric Michael Smith, Melissa Hall Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. " I'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209* (2022).
[62] HE Stanton. 1971. Incidental and intentional learning—one process or two? *Australian Psychologist* 6, 1 (1971), 26–30.
[63] Danny Sullivan. 2018. How Google autocomplete works in Search. *Retrieved November* 22 (2018), 2018.

[64] Danny Sullivan. 2019. How we keep Search relevant and useful. *Google, The Keyword (blog): July* 15 (2019).
[65] Danny Sullivan. 2020. How Google autocomplete predictions are generated. *Retrieved October* 8 (2020), 2020.
[66] Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (2019-11-04), Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett (Eds.). 13209–13220. http://papers.nips.cc/book/advances-in-neural-information-processing-systems-32-2019
[67] Madalina Vlasceanu and David M Amodio. 2022. Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences* 119, 29 (2022), e2204529119.
[68] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (sep 2014), 78–85. https://doi.org/10.1145/2629489
[69] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. *Measuring and Reducing Gendered Correlations in Pre-Trained Models.* http://arxiv.org/abs/2010.06032
[70] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.
[71] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
[72] UN Women. 2013. UN Women ad series reveals widespread sexism. *Un Women* 21 (2013).
[73] Yahoo. 2023. *Yahoo Search Autocomplete Policy.* hhttps://help.yahoo.com/kb/SLN36183.html#/
[74] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA, 2018-12-27) *(AIES '18)*. Association for Computing Machinery, 335–340. https://doi.org/10.1145/3278721.3278779
[75] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).

A ADDITIONAL TABLES

Table 10: Average number of autosuggestions per group (US January 2022)

Category	# groups	Google	Yahoo!	Duck.
Age	9	2.4	6	5.6
Gender/gender identity/gendered	23	3.9	3.4	2.5
Lifestyle/way of living	14	2.1	1.4	0.8
Political/politicised	8	0.4	3.6	0.3
Peoples/ethnicities	32	1	3.2	0.5
Nationalities	47	1.1	1.1.	1.9
Religion	11	0.2	3.8	0
Sexual orientation	7	0	0	0

Table 11: Average number of autosuggestions per group (US August 2022)

Category	# groups	Google	Yahoo!	Duck.
Age	9	2.3	5.3	5.6
Gender/gender identity/gendered	23	4.3	2.8	2.3
Lifestyle/way of living	14	2.2	1.4	0.0
Political/politicised	8	0.4	3.8	0.1
Peoples/ethnicities	32	0.9	2.8	0.4
Nationalities	47	1.3	0.8	1.9
Religion	11	0.5	3.5	0
Sexual orientation	7	0	0	0

Table 12: Examples of synthetic patterns appended to organic search logs

DuckDuckGo	Why are boys so [ugh]
	Why are brothers so [sweaty step brothers, annoying quiz]
	Why are kids so [entitled now]
Yahoo	Why are Americans so [stupid 2020, angry ielts reading]
	Why are boys so [stupid quotes]
	Why are Black Americans so [much bigger than Africans today]
	Why are Muslims so [spoken word]
	How come homeless people are so [sensitive quotes]
	Why are daughters so [mean to their mothers now]
	Why are Russians so [cruel people images]
	Why are kids so [sensitive these days, dumb today]
	Why are teens so [sad derek thomson, sad today]