Article

new media & society
2025, Vol. 27(8) 4865–4890

DOI: 10.1177/14614448241244735
journals.sagepub.com/home/nms
Sage

Influence of hate speech about refugees in search algorithms on political attitudes: An online experiment

Franziska Pradel
Technical University of Munich, Germany; University of Cologne, Germany

Abstract

This article assesses the effects of hate speech compared to positive and neutral content about refugees in search engines on trust and policy preferences through a survey experiment in Germany. The study uncovers that individuals with an extreme-right political ideology become more hostile toward refugees after being exposed to refugee-related hate speech in search queries. Moreover, politically biased search engines erode trust similarly to politicized sources like politicians, and positively and negatively biased content is trusted less than neutral content. However, individuals with a right political ideology trust more hate speech content than individuals with a left-wing ideology. Individuals with right-wing ideology are also almost three times as likely to intend to click on hate speech suggestions compared to left-wing counterparts.

Keywords

Algorithmic bias, attitudes toward refugees, consequences of hate speech, hate speech, polarization, political information-seeking, search engines, trust

It is the 21st century, the digital landscape prompts individuals to seek immediate political information online. However, seeking information online is not immune to political biases; algorithm-driven search engines are generally trusted and heavily relied on—but can recommend politically biased content. This study investigates how individuals, anchored by political ideologies, are causally affected by navigating biased algorithmically generated recommendations in search engines.

Corresponding author:
Franziska Pradel, Department of Governance, TUM School of Social Sciences and Technology, Technical University of Munich, Richard-Wagner-Straße 1, 80333 Munich, Germany.
Email: franziska.pradel@tum.de

Search engines like Google play an essential role when a political topic is salient because they are used extensively, are trusted, and perceived as neutral (Dutton et al., 2017; Pan et al., 2007; Purcell et al., 2012; Schultheiß et al., 2018), but can also reflect biased and derogatory content (Baker and Potts, 2013; Haak and Schaer, 2022; Noble, 2018; Otterbacher et al., 2017; Pradel, 2021). Focusing on search engines is crucial due to their role as gatekeepers to online information (Dutton et al., 2017). Particularly when a political issue is salient in the public, if it gains public attention, people turn to search engines to gain immediate information, and to follow the issue closely.

One issue that has widely received the attention of the public and the media is immigration and refugees since the so-called "refugee crisis" in 2015 in Europe, marked by large refugee flows into Europe and the rise of skepticism toward refugees within public opinion in the aftermath (Dennison and Geddes, 2019; Eberl et al., 2018; Wike et al., 2016). Several studies have shown that group cues and negative sentiments in communication about immigrants both in media outlets (Brader et al., 2008; Czymara and Dochow, 2018; Eberl et al., 2018; Wirz et al., 2018) and from political elites (Newman et al., 2021) are prone to reinforce negative attitudes about refugees and immigration. Refugees have not only been frequently portrayed in the European press with narratives promoting stigmatization, suspicion, hate speech, and hostility (Georgiou and Zaborowski, 2017; Wigger et al., 2021), they are also targeted in more extreme forms like hate speech when it comes to newer forms of mass media like social media, where also organized networks of individuals actively engage in spreading hostility (Gagliardone et al., 2015; Mathew et al., 2019). Especially hate speech has far-reaching negative consequences, as it can discourage political participation (Special Eurobarometer 452, 2016).[1] Research is nonetheless limited when it comes to new technologies and algorithmically curated content—like from search engines—despite their wide use for actively seeking information.

Positioned as the initial suggestions users encounter when seeking online information, search engines' autocompletions play the role of guiding prompts, suggesting completions based on their algorithms. With this in mind, this study uses an online experiment in Germany to examine the causal effects of hate speech (i.e. negative stereotypical), positive and neutral content in search engines' query suggestions on political attitudes and trust, and the role of political ideology. In Germany, the rise of right-wing populism has been observed alongside the increasing importance of anti-immigrant sentiments and polarization in the society (Franzmann et al., 2020), which makes it an interesting case to study.

The study provides evidence of polarization among individuals at the extreme ends of the political spectrum—those with both extreme left and extreme-right ideologies—when they encounter algorithmically recommended hate speech through search engine autocompletion. This hate speech effect appears to influence attitudes primarily at the fringes. For instance, exposure to such content seems to particularly fuel more hostile attitudes toward asylum policy among individuals with an extreme-right political ideology. However, this polarization does not manifest when encountering positive content within search engines, but attitudes converge across the spectrum. The study also shows the substantial trust placed in search engines, though this trust is jeopardized when search query suggestions exhibit pronounced bias, causing erosion in trust for the otherwise highly trusted source.

The study findings reveal that political ideology notably affects trust, as right-wing participants trust such hate speech content more than left-wing participants. Even though the hate speech effects might only last for a short time, those who are strongly interested in the topic would probably actively search for the topic more often. The overall effect might be more lasting for individuals given how often people search for this content. In addition, explorative analyses show that individuals would also react more sensitively to biased political information when seeking it online, depending on their political ideology; that is, people with a right-wing political ideology are almost three times more likely to click on hate-speech suggestions than those with a left-wing political ideology, and they also exhibit more intuitive trust in such suggestions. Therefore, effects may manifest here more strongly.

The significance of biased content about refugees in Germany

Germany offers a particularly good case to investigate the effects of hate speech, positive speech, and neutral speech about refugees in search engines. Since the so-called "refugee crisis," refugees and migration as a general topic have been the focus of public attention (Franzmann et al., 2020). It was the beginning of the rise of the right-wing populist party Alternative for Germany (AfD), which by comparison was strongly opposed to a culture of welcome and found support among parts of the population (Lees, 2018; Mader and Schoen, 2019). An increasing anti-Islamic mood could be observed, and there were public demonstrations by PEGIDA (Patriotic Europeans Against the Islamisation of the West) and AfD against immigration from Syria and other countries (Deutsche Welle, 2017; Lees, 2018).

In Germany and other European countries, refugees got media attention but often negatively, leading to negative attitudes toward immigration, negative stereotypes, and stigmatization (Czymara and Dochow, 2018; Georgiou and Zaborowski, 2017; Wigger et al., 2021). For instance, after the Cologne assaults on New Year's Eve, the immigration coverage has led to increased negative portrayal and criminalization of male migrants (Wigger et al., 2021). Moreover, also political parties increased their attention during this time, with the radical right playing an essential role in shifting attention to this issue, as a study points out (Gessler and Hunger, 2022). The electorate has been rather divided on immigration, that is, the anti-Islamic and anti-migration mood of the electorate opposed other parts that were more pro-immigration (Franzmann et al., 2020), though, it also needs to be noted that the German society is less polarized than the United States (Gidron et al., 2020). The relevance of the topic is particularly evident from the fact that since 2015 it has been mentioned by citizens as "one of the most important problems" in Germany (Forschungsgruppe Wahlen, 2020). Moreover, more than 61% of Germans believe that mainly violent criminals come to Germany (Richter et al., 2023), and 61% of Germans think, according to Pew Research, that refugees will increase the likelihood of terrorism (Wike et al., 2016).

Therefore, biased content about refugees may further affect the political attitudes of German citizens.

Hate speech and its consequences

Prior research supports the assumption that individuals use stereotypes and categorize individuals into different groups to simplify the information processing (Allport, 1958; Tajfel, 1970). Stereotypes are, often automatic, generalizations and associations with people based solely on their belonging to a group, for instance, characteristics, character traits, and behavior (Fiske, 1998; Katz and Braly, 1935). This involves assigning positive associations and qualities to the in-group, which refers to members of the group with which a person identifies, and negative associations to the out-group (Fiske, 1998). Through generalizations to the whole group, stereotypes increase between-group differences and decrease within-group differences (Fiske, 1998: 357). This stereotyping can lead to intergroup bias and discrimination as individuals gain self-esteem by favoring people from their social group (in-group), that is, people with similar characteristics to themselves, and by derogating people outside this group (out-group) (Billig and Tajfel, 1973; Tajfel, 1970; Tajfel and Turner, 1979). As a result, individuals tend to see others of the group they identify with as positive and those outside the group as negative (Billig and Tajfel, 1973). This study looks into the effects of hate speech toward refugees—a form of derogation of an out-group—in the new media platform search engines. Hate speech and organized hate are increasingly observable in the online world; they derive their strength from the fact that hate is the most potent negative emotion for mobilizing people for derogation and violence toward an out-group (Fischer et al., 2018).

Hate speech has been commonly defined in the literature, in a broader sense, as hostility or promoting violence toward a social group based on characteristics such as race, ethnicity, gender, and religion (Gagliardone et al., 2015; Koltsova et al., 2017; Mathew et al., 2019; Waltman and Mattheis, 2017). However, definitions and conceptualizations vary when it comes to specifics. It has been highlighted to be a challenging task in computer science studies detecting hate speech in large-scale data with the help of human-annotated data, especially when providing only broad and vague evaluation categories, as the assessment can be subjective (Koltsova et al., 2017). Research specifically points out that the understanding of hate speech varies depending on its situatedness (Udupa and Pohjonen, 2019), context, and targets, leading to different considerations about what is considered harmful and merits moderation (Pradel et al., 2024).

In addition to hate speech, other important concepts have been introduced in the literature that cover speech that is potentially dangerous to societies. These conceptualizations, in contrast to the conceptualization of hate speech also used in this study, are more comprehensive as they take into account broader contexts or varieties of speech. Udupa and Pohjonen (2019) highlight the importance of the situatedness of online speech varying with different milieus globally, thereby also considering the everyday-life granularity of online practices and the political, economic, cultural, and historical contexts. In comparison, other work has emphasized the consequences of speech in their conceptualization of harmful speech. According to Buyse (2014), for instance, fear speech refers to the potential of speech to incite fear against another group, which makes it more likely to accept violence toward the respective group and thereby, as argued by the author, more relevant than general hate speech for the risk of incitement of violence. Moreover, hate speech can be distinguished from incivility, which is defined in a wider sense as all uncivil behavior, including milder forms such as name-calling and more extreme forms such as hate speech against a social group (Papacharissi, 2004; Ziegele et al., 2018).

While I acknowledge the importance of considering context and situatedness like proposed in the extreme speech framework (Udupa and Pohjonen, 2019) for (future) comparative work on the subject, in this study, I follow a practical approach. Specifically, I focus on derogatory, negative stereotypical content targeted at a protective group (i.e. immigrants), which is in line with major platforms' definition to tackle hate speech. For instance, Meta considers harmful stereotypes and generalizations as hate speech, such as generalizations as criminals or generalizations about inferiority like the mental state, which would also categorize the subsequently presented experimental stimuli as hate speech.[2]

Research on the effects of hate speech and incivility toward minorities in a society shows that hate speech attitudes can reinforce negative prejudices, aggressiveness and violence toward the minority group (Perry, 2001; Waltman and Haas, 2011; Waltman and Mattheis, 2017). On one hand, it can change the emotional status of recipients, for example, by leading to more enthusiasm (Kosmidis and Theocharis, 2020) or entertainment (Nikolaev et al., 2023). On the other hand, it can have devastating consequences for the minority group and the society because hate speech can lead to various negative consequences, such as negative psychological outcomes (e.g. depression (Bilewicz and Soral, 2020), lower self-esteem (McCoy and Major, 2003), to work-related problems (Klaßen and Geschke, 2019), political consequences (e.g. lower participation in debates; Special Eurobarometer 452, 2016), political intolerance (Halperin et al., 2009), and social consequences (e.g. lower willingness to donate to a refugee aid organization; Ziegele et al., 2018). By promoting hostility against social groups, hate speech can contribute to a climate that may ultimately facilitate violent acts (Perry, 2001; Waltman and Mattheis, 2017). The right-wing extremist act of terror in Hanau in Germany (Hoffman et al., 2020) and the Pittsburgh synagogue shooting in the United States (Mathew et al., 2019; McIlroy-Young and Anderson, 2019) are examples where hate speech may have played a role—among various complex factors—contributing to these tragic events. As hate speech becomes increasingly prominent on online platforms, especially in recent years, the problem that hateful content can spread faster than ordinary content is also becoming apparent (Mathew et al., 2019).

Negative stereotypes about groups also play an important role in hate speech. Research found that priming of stereotypes or racial attitudes, even if only subtly, can affect political attitudes. It can lead to decreasing support for political candidates (Valentino, 1999) and increased criticism about them (Pyszczynski et al., 2010). Moreover, priming stereotypes can increase intergroup conflict (Hsueh et al., 2015) and lead to more support for certain policies like punitive crime policy agenda (Gilliam and Iyengar, 2000).

Hate speech and stereotypical information in search engines

Search engines are a frequently used medium. As representative studies show, about 78% of Germans use search engines at least on a weekly basis (Beisch and Koch, 2022).[3] A study from 2017 has found comparably high values in the frequency of search engine use in several countries (Dutton et al., 2017).[4] However, search engines have a vast potential to confront users with prejudices and stereotypes. For example, Google's search autocompletions (also called search suggestions, search predictions)—information that pops up when searching for a term in the search bar—are offered based on what prior users have searched for. Research on intergroup relations has established that people tend to categorize others into groups and cling to stereotypes (Allport, 1958; Cuddy et al., 2009; Tajfel and Turner, 1979), how individuals search for information and, thus, create content about social groups is likely to be influenced by stereotypes and prejudices. Particularly, research found that Google contains several sexist and racist search results, images, image-labeling, and map locations (Noble, 2018). For example, racial slurs redirected users to the White House during Obama's presidency and Google Photos categorized a photo of Blacks as "gorillas." Other research indicates that search engines contained anti-Semitic (Bar-Ilan, 2006), gender-stereotypical (Otterbacher et al., 2017), and stereotypical and/or negative content for certain social groups (Baker and Potts, 2013), as well as biased information about politicians (Haak and Schaer, 2022; Pradel, 2021).

Crucially, users having more stereotypes (i.e. gender stereotypes) about a group are also less likely to notice these biases (Otterbacher et al., 2018) and they even perceived the reality more stereotypically (Kay et al., 2015), when being exposed to such content. Research also indicates that the users' behavior in search engines is also crucial for the exposure to potentially politically biased information. A search engine may expose them to diverse information when neutrally formulated search queries are used and a variety of the search results is considered (Steiner et al., 2022). Importantly, however, research suggests that users formulate different queries depending on their ideology, a process in which they may expose themselves to information that confirms rather than challenges their opinion (Van Hoof et al., 2022).

As gatekeepers of information, search engines can play a crucial part when they confront individuals with hate speech about minorities since research showed that negative stimuli attract and affect individuals to a large extent (Soroka and McAdams, 2015). Although Internet users resort to search engines frequently, prior research showed the existence of biases toward minority groups and suggested users' attraction to it; it remains unknown how they affect political attitudes. This study focuses on this gap and compares the effects of hate speech to the effects of positive and neutral content about refugees as well as to a control condition.

Expected effects of search engines on political attitudes

Expected effects of hate speech on political attitudes

Building on research on intergroup conflict (Tajfel, 1970; Tajfel and Turner, 1979) and motivated reasoning (Kunda, 1990), the study examines how content with derogating—compared to positive and neutral—content about minority groups affects preferences for political policies related to refugees (i.e. immigration and asylum policies).

The social identity theory (Tajfel, 1970; Tajfel and Turner, 1979) was proposed to explain intergroup behavior. According to the theory, individuals belong to different social groups, and individuals categorize themselves and are also categorized by others into social groups and tend to distance themselves from out-group members. One proposition is that individuals favor members of their social group who share similar characteristics, such as ethnicity and gender (in-group favoritism), and engage in derogating others (out-group derogation).

For instance, a significant group is *refugees in European countries*, particularly relevant due to the public attention they have received and the increased unease surrounding immigration following the so-called "refugee crisis," predominantly among those with pre-existing concerns about immigration (Dennison and Geddes, 2019). Relatedly, the portrayal of immigrants in media content and especially a negative sentiment can lead to negative attitudes toward immigrants, including migrants and refugees (Brader et al., 2008; Czymara and Dochow, 2018; Wirz et al., 2018). Moreover, research showed that hate speech could reinforce stereotypes, increase aggressiveness, and even lead to violence (Perry, 2001; Waltman and Haas, 2011; Waltman and Mattheis, 2017). It is reasonable that individuals strengthen their attitudes when they are confronted with hate speech in search engines, as it is specific content that is automatically recommended to them and implies the interest of previous search engine users on this issue, which may make them feel validated in their basic negative sentiment toward refugees (Baker and Potts, 2013).

Motivated reasoning (Kunda, 1990) and the related confirmation bias (Oswald and Grosjean, 2004; Wason, 1968) describe the cognitive process wherein individuals tend to approve and accept information that is aligned with their attitudes. It may also provide a relevant theoretical framework for explaining the effects of hate speech on political attitudes in this context. Specifically, when search engines recommend their users hate speech content about refugees, these users may be more inclined to accept the content and be influenced by it when it strongly aligns with their pre-existing negative sentiments toward refugees. Ultimately, this may lead to the assimilation of the recommended negative content and reinforcement of their attitudes related to refugees. In light of these considerations, it is reasonable that individuals will become more critical toward refugees in terms of immigration policy and asylum policy after being exposed to hate speech.

> *H1*. Exposing individuals to hate speech in search autocompletions will make them more critical toward refugees.

However, it is also plausible that hate speech is not always processed in the same way, but it needs to be considered that individuals' political ideology plays an essential role in processing algorithmically recommended political information. Reiterating the importance of motivated reasoning (Kunda, 1990) and confirmation bias (Oswald and Grosjean, 2004; Wason, 1968), it is apparent that people tend to favor and accept information the more it reinforces their viewpoints. Given this consideration, it is important to highlight that individuals with a right and less progressive political ideology tend to have more restrictive attitudes toward refugees and immigration policies than those with a leftist political ideology (Liebe et al., 2018). Research also showed that the effects and perception of incivility could vary with individuals' political ideology and partisanship (e.g. Costello et al., 2019; Kosmidis and Theocharis, 2020). In other words, it seems reasonable to anticipate disparate reactions to hate speech in search engines based on individuals' political ideologies:

> *H2*. Individuals with a right-wing political ideology will become more critical toward refugees than those with a left-wing political ideology when being exposed to hate speech in search autocompletions.

Expected effects of positive content on political attitudes

Nevertheless, positive expressions related to social groups could be harmful too when they provoke reactance (Brehm, 1966; Burgoon et al., 2002 see also backfire, boomerang, or backlash effect, for example, Swire-Thompson et al., 2020; Nyhan and Reifler, 2010). There are mixed findings, however, on whether a backfire effect exists. Some studies found evidence of a backlash after threats to beliefs (e.g. Anduiza and Rico, 2022; Hart and Nisbet, 2012; Nyhan and Reifler, 2010) among subgroups like strong Republicans who amplified their attitudes (e.g. Hart and Nisbet, 2012), while others did not find an impact (e.g. Guess and Coppock, 2020; Haglin, 2017; Wood and Porter, 2019), calling for further research.

Similarly, stemming from theoretical arguments as outlined in the cognitive dissonance theory (Festinger, 1957), deviating worldviews from the own are causing discomfort to individuals and people are eager to avoid uncertainty, stress, or anxiety arising from such inconsitencies with their beliefs and attitudes (Festinger, 1957). As outlined by Festinger, reactions to dissonance are powerful and appear across a wide range of contexts, including reactions to deviating political attitudes. At the same time, stress, anxieties, and threats, that may also arise through inconsistencies, can provoke coping mechanisms with people reinforcing their pre-existing meaning frames and attitudes (e.g. Brandt and Crawford, 2020; Proulx and Major, 2013; Rovenpor et al., 2016). But research is needed to determine whether it also applies to algorithmically recommended content about refugees that is provided during the search process, like in search suggestions. Worldviews deviating from one's own—for example, the worldview full of conflicting positive content one might see in search suggestions—can be considered a threat that may provoke coping strategies.

However, should we expect similar effects like in previous research on more conventional (social) media settings? Technology is generally trusted and often perceived as neutral and unbiased. This perception may stem from the absence of an individual, a human, as the central sender; instead, it originates from a complex software system—a machine designed for retrieving information. Search engines, in particular, may influence fundamentally different in that they essentially recommend information to their users and are perceived as neutral. The reason is that people may use different heuristics for algorithm-curated content, like topic suggestions by search engines, and trust them more than individuals because they are more aware of humans' potential bias and emotions, knowing they often follow their own agenda. In contrast, "machine heuristics" suggest the instinctive belief that "machines are more objective than humans, can perform tasks with greater precision" (Sundar and Kim, 2019: 2), and the influence of their recommendations can even surpass social influences according to recent arguments and empirical findings (Bogert et al., 2021; Sundar and Kim, 2019).

As outlined earlier (see Section "The significance of biased content about refugees in Germany"), the majority in the German society is rather skeptical and critical toward refugees. Positive content about refugees can provoke stress for people who have a rather critical attitude toward refugees that they need to cope with, which leads to the following hypothesis:

> *H3*. Exposing individuals to positive content in search autocompletions will make them more critical toward refugees.

Once again, political ideology may be a crucial moderator when it comes to positive content about refugees. Specifically, positive content about a minority group may provoke more stress, and thus, more need for coping strategies by becoming more critical toward refugees for individuals self-identifying at the "right" of the political spectrum—leading to the hypothesis:

> *H4*. Individuals with a right-wing political ideology will become more critical toward refugees than those with a left-wing political ideology when being exposed to positive content in search autocompletions.

Expected effects of search engines as a source on trust

Besides the political ideology, the source that provides individuals with hate speech may play an essential role in information processing. It may be central that search engines are generally perceived as neutral and that their information is trusted. By comparing hate speech coming from search engines to another source frequently covering the topic of refugees (i.e. a politician), the relative importance of web technologies compared to humans as sources of hate speech can be estimated. A political actor may be perceived as less objective and less trustworthy than search engines as sources of communication. Other research showed that although individuals trust search engine results less than some years ago, users still tend to trust them and rely intuitively on the search results' outputs and ranking (Joachims et al., 2017; Schultheiß et al., 2018; Schultheiß and Lewandowski, 2023).

According to a study by the Pew Research Center, most Americans perceive them as unbiased and fair sources of information (66 vs 20%) and that all or almost all (28%) or most (45%) of their information is trustworthy and accurate (Purcell et al., 2012). In Germany, more people perceive Google as correct and trustworthy (46 vs 6%), fair and unbiased (34 vs17 %) versus the opposites, according to a representative study (Schultheiß and Lewandowski, 2023). Another study with cross-country data revealed similar trust ratings in search engines as information sources among Internet users in the US and European countries (total 52 vs 8%) like Germany (44 vs 11%; Dutton et al., 2017:

42–43). However, when asked about trust in news in search engines, only 21% (vs 25%) of representative German survey participants indicated trust, but as well as in social media platforms (10 vs 45%; Jackob et al., 2023: 51).

According to these numbers, there is a rather general trust in search engines in the public, and people may instinctively believe in their objectivity and reliability. They may even trust biased content more in search engines than other sources, particularly when compared to biased content from political figures that are often associated with inherent biases, political agendas, and less objectivity.

Thus, individuals may intuitively perceive search engines as more neutral than politicians and trust their content. This leads to the following two hypotheses:

> *H5*. Biased content provided by a search engine will be more trusted than the same content provided by a politicized source, that is, a politician.
>
> *H6*. A search engine will be more trusted than a politicized source, that is, a politician.

Current study for investigating expected effects

To investigate these expectations, this study uses an online survey experiment in Germany that primes participants in the experimental group with biased (hate speech and positive) search suggestions and neutral suggestions while comparing them with content that is not refugee-related (control group). Within this study, a particular focus is the analysis of the causal effects of such biased content on policy preferences related to refugees and trust in search engines in general and their content. Moreover, at the end of the study, explorative analyses will also uncover the click intentions of participants differentiated by their political ideology.

Design, procedure, and measures

Before running the experiment, the ethic commission of the Faculty of Management, Economics and Social Sciences at the University of Cologne approved the study (approval number: 19020FP), which has been pre-registered on EGAP.[5] All presented expressions of the prime stimuli were validated by human coders who rated a set of randomly presented expressions into either a neutral, negative, or positive category.

Participants

In total, 1200 participants, including 607 women and 593 men, with a minimum age of 18 years have been recruited by an online panel (i.e. Lucid, collected in April 2020). The platform provides reliable and valid data for online experiments when using attention checks (Coppock and McClellan, 2019). Quotas have been applied so that the recruited participants are representative of German citizens in terms of gender, age, and education. In addition, the online experiment included an attention test by asking participants to click on a particular category to have a test of data quality. Participants who did not

Table 1. Experimental groups: hate speech, neutral, and positive content provided by a search engine (translated).

Hate speech	Neutral tone	Positive tone
Refugees are criminal	Refugees are in Germany	Refugees are peaceful
Refugees are currently in the debate	Refugees are currently in the debate	Refugees are currently in the debate
Refugees are a danger	Refugees are a group of people	Refugees are a cultural enrichment
Refugees are less intelligent	Refugees are diverse	Refugees are intelligent
Refugees are not assimilable	Refugees are in Europe	Refugees are assimilable

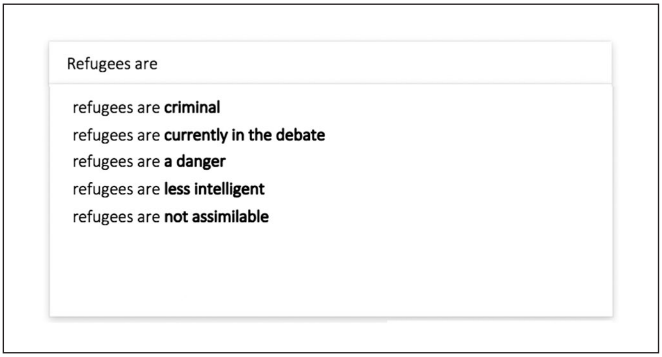
answer correctly to the attention test have been excluded from the data analyses. Balance tests indicate that the randomization worked by suggesting balanced treatment groups ($p > .05$).

Experimental design and measures

The experiment used a 4 × 2 (source: search engine vs politician × tone: control content vs neutral vs positive vs negative speech about refugees) between-subjects design. After providing consent to participate and answering questions about their sociodemographic information, they answered questions about their political ideology (measured with the left-right self-placement, see Breyer, 2015). The respondents were then randomly assigned to one of the eight experimental groups with content related to a minority group (i.e. refugees).

To prime the tone of the information about refugees, they were randomly assigned to either negative, neutral, or positive expressions about refugees or a control group without any expressions about refugees. The negative treatment group contained expressions such as "Refugees are a danger." The positive group contained expressions like "Refugees are peaceful," and the neutral one, expressions such as "Refugees are in Germany." All treatment groups included one expression being more neutral "Refugees are currently in the debate." This has been done to make the search suggestions appear more natural since the suggestions are mixed and rarely solely negative or positive. The translated expressions are shown in Table 1 (originally in German). The participants were asked to imagine that a search engine—versus a politician in the other experimental groups—provided content related to a political topic. Figure 1 shows exemplarily the stimuli of hate speech in a search engine presented to one experimental group. Half of the participants were assigned to the search engine source, the primary focus of this study, while the other half were assigned to a politician in order to compare trust levels when content is biased, juxtaposing the search engine source against a generally less trusted and more politicized source.

Then, the participants answered questions covering attitudes toward immigration policies (Jowell et al., 2007) like "What about immigrants coming from the poorer countries within Europe? Should Germany allow . . ." with a response scale ranging from "allow many to come and live here" (1) to "do not allow anyone" (4) and asylum policies



Refugees are

refugees are **criminal**
refugees are **currently in the debate**
refugees are **a danger**
refugees are **less intelligent**
refugees are **not assimilable**

Figure 1. Experimental group: hate speech provided by a search engine (translated from German).

like "The state should be generous when considering asylum applications." with a scale ranging from "strongly agree" (1) to "strongly disagree" (5) (Prinz and Glöckner-Rist, 2009). Items were summarized to mean indices where a higher mean index score indicates more negative attitudes toward immigration ($\alpha = 0.9$) and asylum policy ($\alpha = 0.84$)—after confirming that the items measure the same construct with exploratory factor analysis (eigenvalue > 1, and suggested by screen plot). Next, they responded to questions about whether they trust the content and source. Finally, the click behavior of the participants has also been measured at the end of the survey experiment by displaying a randomized list of positive, neutral, and negative search suggestions and asking which suggestion they would like to click on to see the corresponding search results.

Besides validating the treatment groups before the experiment by human coders who categorized all single expressions of the treatment groups (randomized) into either negative, positive or neutral, participants also rated on a scale how positive or negative the information about refugees was at the end of the online experiment. These tests also confirmed that the manipulation worked effectively, as the positive manipulations were rated as significantly more positive ($p < 0.05$) and the negative ones as more negative ($p < 0.001$) than neutral experimental conditions. Finally, all respondents were debriefed after the survey experiment.[6]

Results

Attitudes toward immigration and asylum policy

Figure 2 displays results from multiple linear regressions that analyze whether hate speech has any direct effects on attitudes about immigration and asylum policy. As expected, it becomes apparent that individuals who have a right-wing political ideology show significantly more restrictive attitudes toward immigration and asylum policy. There is no effect of hate speech on immigration and asylum policy attitudes if looking only at all individuals (including moderates) with right-wing and left-wing ideologies,

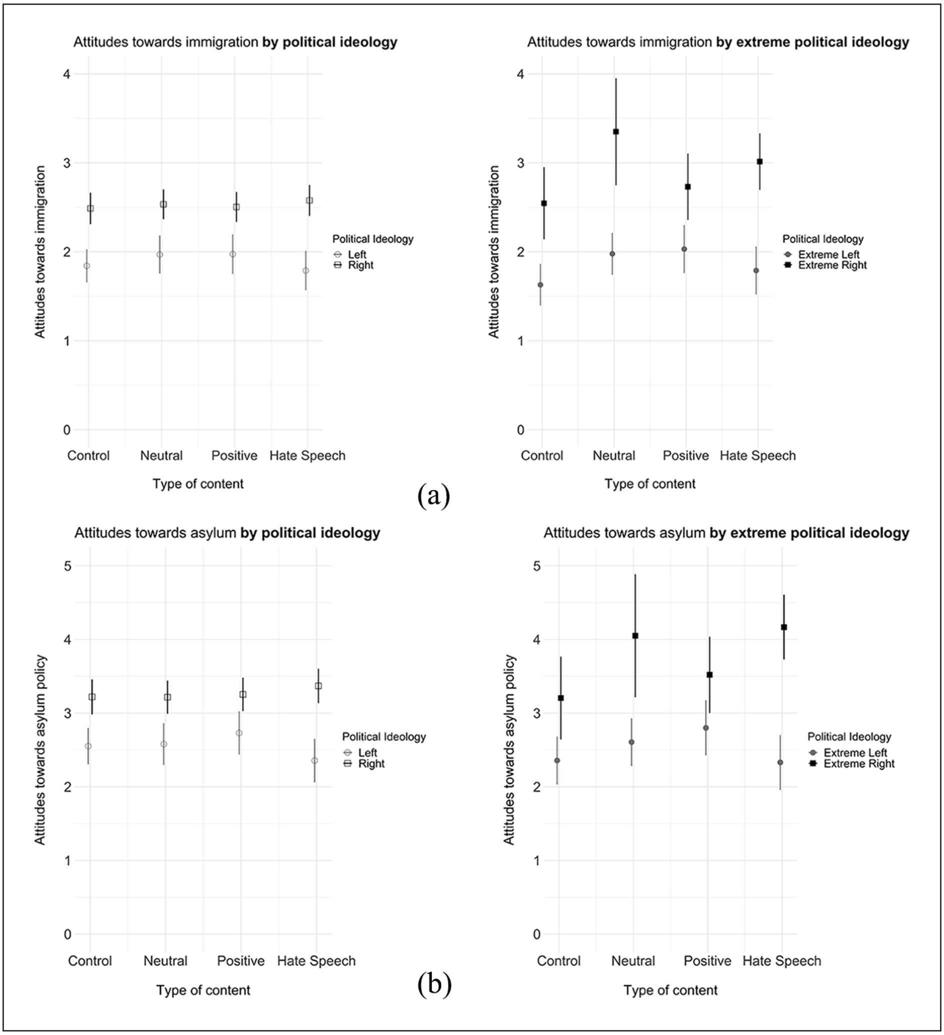


Figure 2. Predicted values and 95% confidence intervals for hate speech, positive, neutral content, and control on attitudes toward immigration (see also Table S3 in the SI for the multiple linear regression results): (a) Attitudes toward immigration policy. (b) Attitudes toward asylum policy.

however, individuals with an extreme-right ideology, that is, those who position themselves at the ideological margins with a self-placement score of 8 or higher on the 10-point scale, where higher values indicate a more right-leaning ideology—become significantly more critical regarding asylum policy and immigration policy following exposure to hate speech. Although, for immigration policy, the difference is only marginally significant with a two-sided *t*-test ($p = 0.074$).[7]

Critically, the distance between individuals with left and right ideologies is larger when all are confronted with hateful search suggestions about refugees as shown in Figure 2. The polarization in immigration and asylum policy attitudes is most pronounced between individuals who are strongly left-leaning (self-placement score ⩽ 3 on the ideological 10-point scale) and those at the far-right end (self-placement score ⩾ 8 on the ideological 10-point scale, see the right panel of Figure 2). When exposed to hate speech, individuals with an extreme left ideology show less stringent attitudes, while individuals with a more extreme-right ideology adopt more restrictive asylum policies. Notably, the distance between the extreme left and right is more pronounced in this context than when comparing the left–right distance among participants exposed to positive content about refugees or the control group.

Crucially, also neutral suggestions about refugees—including statements such as that refugees are in Germany—make extreme-right individuals more restrictive toward immigration and asylum policy as to when they are confronted with positive or no content about refugees.

Trust in the content and search engines as a source

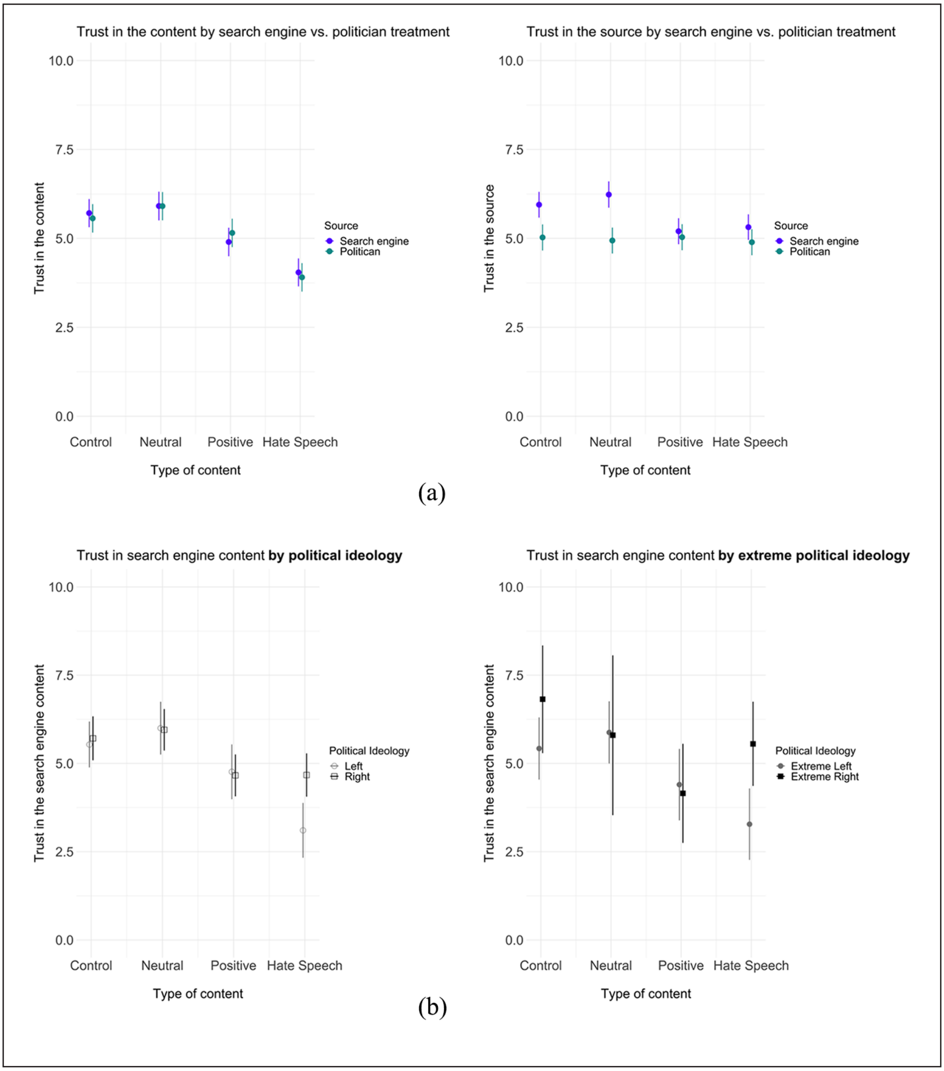
Next, Figure 3 displays predicted values based on multiple linear regressions analyzing whether the participants trust more the content provided by a search engine or the content provided by a politician, the latter being a popular source and typically perceived as politicized (see upper panel of the Figure). Overall, the analyses revealed that participants generally trust most the content provided by search engines and politicians when it contains neutral content about refugees (see the first panel). However, it also shows that they trust the content significantly less if it contains positive content about refugees—and trust least and rather mistrust negative content about refugees.

Remarkably, as expected, individuals seem to trust search engines to a large extent, and the general trust in the source—when having no positive or negative sentiment—is much higher for the search engine than for politicians (see the second panel of Figure 3). If the information has a strong political bias, either positive or negative, toward refugees, however, individuals perceive it as politicized and their content as less credible. In this case, general trust in search engines erodes to a similar degree as the one in politicians.

Importantly, the lower panel of Figure 3 shows trust in different types of search engine content also by political ideologies. It shows that when individuals are exposed to hate speech, those with a right and even more those with an extreme-right political ideology trust more in the content than those with a left or extreme left ideology. Among the extreme right, trust levels are similarly high for both hate speech and neutral content. Thus, political ideology is vital in how individuals trust toxic content. In contrast, while positive content about refugees is trusted less than neutral content, there is no trust gap between the left and the right.

Clicks on hate speech and positive content

Finally, to get a glimpse at how crucial political group identities are entangled with political online behavior, a single logistic regression explored participants' intended



(a)

(b)

Figure 3. Predicted values and 95% confidence intervals based on OLS for hate speech, positive, neutral content, and control on trust in content and general trust in the source for the experimental treatments search engine versus politician (see also Table S4 (panel a) and Table S5 (panel b) in the SI for the multiple linear regression results): (a) Trust in the content and in the source by search engine versus politicized source. (b) Trust in the search engine content by political ideology.

clicks on hate speech search suggestions about refugees compared to neutral and positive suggestions by ideology (see Figure 4). As previously outlined, participants were shown here a randomized list of hate speech, positive, and neutral search suggestions and were asked to indicate which suggestion they would like to click on to see the ensuing search

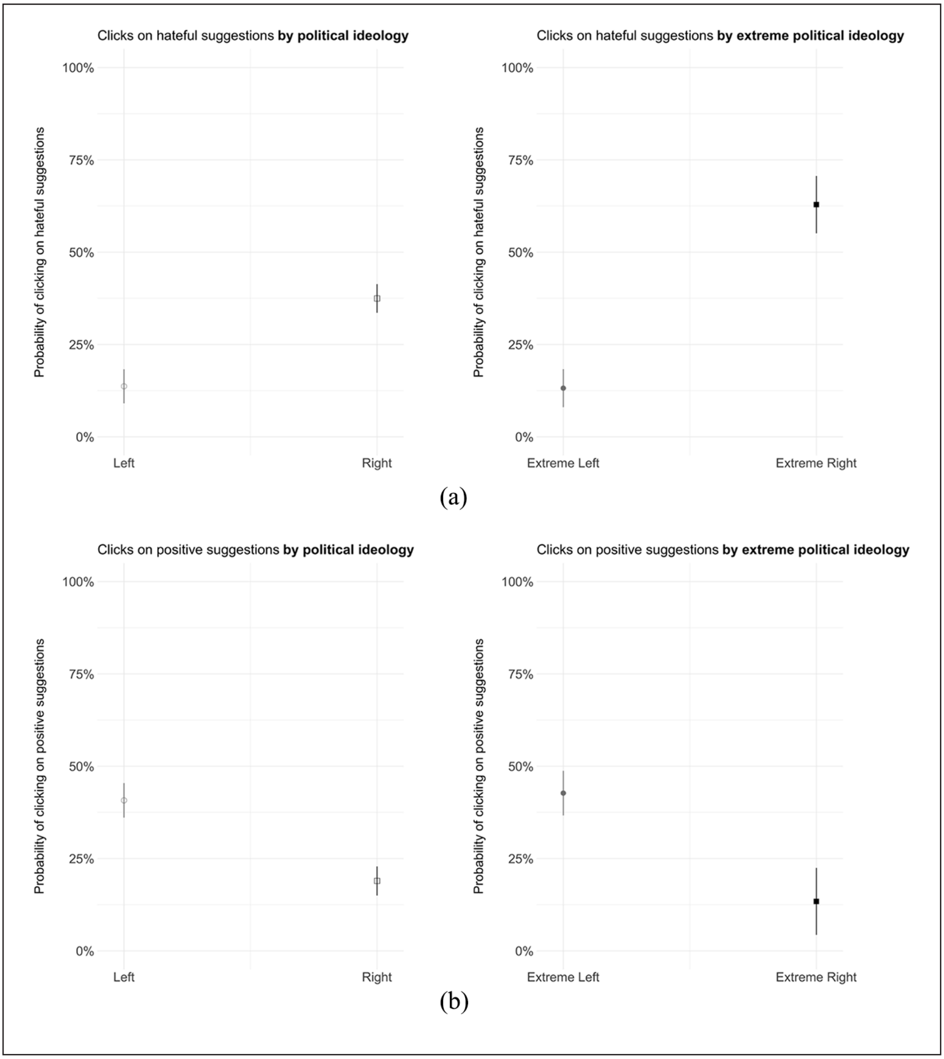


Figure 4. Predicted probabilities and 95% confidence intervals based on a logistic regression of political ideology and extreme political ideology on the intention to click on suggestions with hate speech versus suggestions with positive or neutral content (a) as well as on the intention to click on suggestions with positive speech versus suggestions with hate speech or neutral content (b) (see also Table S6 in the SI for the logistic regression results): (a) Intended clicks on hate speech by political ideology. (b) Intended clicks on positive suggestions by political ideology.

results. While this analysis was not the main focus of the study and only gives insights about correlations, it still may provide important insights into the potential link between participants' political ideologies and their engagement with toxic online content that could be further probed in future research.

The results indicate that individuals' political ideology significantly predicts the likelihood of clicking on a hate speech search suggestion. Having a right-wing political ideology versus having a left-wing political ideology significantly increases the log odds of clicking on a search suggestion with hate speech. The probability of clicking on hate speech content is about three times as high for the users with a right-wing ideology (37%) than for those with a left one (14%; see Figure 4). Moreover, the predicted probability of an individual with an extreme-right political ideology indicating a click preference to hate speech content stands at 63%, which makes them about five times more likely to do so than those with (extreme) left-wing ideologies (13%). Similarly, the lower part of the figure also shows click intentions for positive suggestions and illustrates that (extreme) left-leaning individuals are more likely to indicate click intentions for such suggestions than right-leaning individuals, supporting the arguments of motivated reasoning and cognitive dissonance.

Discussion and conclusion

The results show that those with an extreme-right political ideology reinforced their hostile attitude toward refugees and supportive actions after being exposed to hate speech in search engines, but there was no substantial effect among the average participants (including political moderates). This could indicate that hate speech content about refugees triggers these individuals' negative opinions that are aligned with their political group identity (e.g. criminal; asylum fraud), which is generally reflected in reinforced restrictive attitudes toward asylum policy.

These results also align with studies that have demonstrated that motivated reasoning effects are commonly more pronounced among those with strong beliefs on an issue or when the issue is strongly tied to their identity than those with moderate attitudes. As highlighted by Chong and Druckman (2007: 112, 120), individuals with strong attitudes and values are more prone to engage in motivated reasoning, making them more likely to interpret new information in ways that reinforce their beliefs and resist disconfirming information (see also arguments by Avdagic and Savage, 2021). This phenomenon gains additional support from Avdagic and Savage (2021), who underscore that individuals harboring entrenched anti-welfare and anti-immigrant attitudes manifest an elevated susceptibility to negative framing, thus amplifying their negative perceptions and responses, a finding that resonates with this study results on the importance of extreme political ideology for engaging with the harmful online content. Similarly, this is also in line with studies on backfire effects that have "almost exclusively been found in either political or attitudinal subgroups" (Swire-Thompson et al., 2020), for example, for those with strong beliefs on the issue or when it is strongly connected with the identity (Flynn et al., 2017; Lewandowsky et al., 2012; Nyhan and Reifler, 2010). Notably, in this study, motivated reasoning effects are mostly predominant among the extreme right but less among the extreme left. A reasonable explanation could be that immigration and asylum issues have way more importance to the former, and they have stronger pre-existing negative attitudes toward refugees that are triggered by recommended search suggestions.

Moreover, it may also be the case that participants infer public opinion from the autocomplete suggestions of search engines knowing or assuming that the autcomplete function also shows what previous users have searched for. This resonates also with the persuasive press inference model (Gunther, 1998) and the spiral of silence theory (Noelle-Neumann, 1993). According to these theories, mass media content like newspaper articles can affect individuals' perception of public opinion. This effect is also shown more recently with social media content (Neubaum and Krämer, 2017). In the SI, exploratory analyses look into this question and discuss the findings in more detail. Solely hate speech seems to make people with an extreme-right ideology infer a more hostile public opinion toward refugees when exposed to hate speech, while extreme left individuals rather tend to perceive it as less hostile as compared to when exposed to other content. This may help to explain why predominantly right-extreme individuals seem to become more hostile toward refugees, as they may normalize and legitimize such language, while individuals with more moderate viewpoints show lower susceptibility to these effects.

As shown by the study, trust in search engines is very high and is generally higher than in politicians, who are popular communication sources when it comes to refugees and immigration and asylum policy. The search engine, however, as shown here, is perceived as politicalized when the content is negatively or positively biased. In these cases, general trust declines to a level comparable to general trust in politicians, and trust in the content erodes. This finding also contributes to the existing body of research on trust in search engines (Pan et al., 2007; Schultheiß et al., 2018; Schultheiß and Lewandowski, 2023) and extends our understanding of how trust shifts in case a search engine's algorithm recommends mostly biased content. However, individuals with a right-wing ideology have significantly more trust in hate speech content than those with a left-wing ideology. This finding again aligns with theoretical expectations of motivated reasoning, anticipating that those who hold stronger critical sentiments toward refugees would be more inclined to perceive negative information that reinforces their critical beliefs as being more credible than individuals with more progressive attitudes.

As an explorative analysis in this study revealed, political ideologies also play a significant role in how individuals interact with hatred and positive expressions about refugees. Individuals with a right-wing political ideology were about three times (37%) more likely to click on hate speech than individuals with a left-wing ideology, and individuals with an extreme political ideology were about five times more likely (63%) than those with a left or extreme left political ideology. A further test of click intention by participants' party preferences (i.e. the party they would like to vote for if there were an election next Sunday) confirms these findings by showing that voters of the radical right parties AfD (60% predicted probability) and NPD (100%) predominantly indicated a preference for seeing search results of a search suggestion with hate speech (see Table S7 and Figure S4 in the SI).

One problematic finding is that, in particular, individuals with an extreme-right political ideology are becoming more hostile in their attitudes when being exposed to hate speech. However, looking at the findings from a positive angle, two points stand out: Hate speech in search query autocompletion had minimal impact on the majority of users, and the majority demonstrated limited interest in clicking on hate speech about refugees, especially those categorized as moderate individuals. Furthermore, for the right-extreme group affected by this, the automatically suggested information is more likely to exert a short term rather than a lasting effect. If susceptible groups actively expose themselves to such toxic content, as the research finding suggests, or if the search engine algorithms predominantly show them similar toxic (news) content after clicking on hate speech content, effects may get pronounced. Search engines may be a gatekeeper for hate speech that some groups are more likely to endorse. Thus, the effects could be amplified in the long term through information repetition (Fazio et al., 2015; Swire-Thompson et al., 2020). More research on long-term effects is needed.

This study focused on the effect of hate speech in search suggestions in a specific context: Germany in a time of many citizens expressing animosity or, at a minimum, skepticism. While this work does not suggest that the findings are generalizable to other societies that are less xenophobic, the findings give novel insights into how individuals react to hate speech against a vulnerable group during intergroup conflict and into belief-confirming click behavior. Thus, the findings may also be relevant to societies with ongoing intergroup conflict and animosity.

Moreover, while I acknowledge the increased participation of search engine providers in content moderation, it is crucial to note the evolving landscape of online content. Search engines like Google may maintain higher thresholds for hate speech or pornography than in the past. Examples of derogatory content, such as hate speech, excessive pornographic and sexist content when searching for Black women, or racist slurs that happened in the past, appear less frequently when searching again for the same groups today (Noble, 2018), potentially due to search engine providers' alert to some examples through public attention and critiques followed by increased content moderation efforts and filtering. However, as highlighted by Noble (2018: 11), new instances of stereotypical or derogatory content can emerge.[8] The SI shows past examples of derogatory content and recent examples of biased and negative content that serve as evidence that novel forms of derogatory content can arise, for instance, the disproportionate focus on criminality associated with Syrians in search engines or suggestions that lead to right-wing extremist songs that incite hatred. Similarly, large language models, such as ChatGPT and Aleph Alpha have recently received attention because of hate speech including the praising of Hitler and exhibited derogatory and biased content (Gross, 2023; Lindern, 2023; Tagesspiegel, 2023 see SI for more insights).

It needs to be acknowledged, though, that social media is probably a likelier source of random exposure to hate speech in individuals' everyday lives. However, a systematic analysis of the prevalence of search engines' algorithm-curated content, including search suggestions, search results, recommended news articles, and other content in different languages and regions, remains an important avenue for future research, particularly because moderation and regulation vary across countries. Overall, the study reveals the need for continued and proactive content moderation of algorithmic-curated content across various platforms and tools used for information-seeking, as well as attention to the varying susceptibility among users.

To conclude, search engines, their autocomplete suggestions, which are investigated here in this study, as well as their other algorithm-curated content, remain an underresearched topic, although they are part of our everyday life and are the most used platform to search for information. Particularly, hate speech recommended by the popular online information gatekeeper may induce negative consequences for society because, as shown here, it could nourish polarization of society and extremism. In view of the hate-enhancing potential of hate speech and the high levels of xenophobia, hostility, and aggression against Muslims and migrants (Decker and Brähler, 2018)—and the potential digitization and digital technologies contributing to anti-democratic attitudes and behavior, the need to counteract information structures that promote xenophobia is once again becoming apparent. Thus, the study findings highlight the importance of filtering toxic online content like hate speech in search suggestions.

Acknowledgements

The author's work benefited from detailed comments on earlier versions of the article. She is very grateful to Sven-Oliver Proksch, Bruno Castanho Silva, Jens Wäckerle, Lea Kaftan, Ayjeren Rozyjumayeva, Danielle Pullan, Lennart Schürmann, Jan Schwalbach, Sarah Berens, Yannis Theocharis, and anonymous reviewers for feedback. In addition, she would like to express her appreciation to the Digital Society research program and its members for engaging in fruitful research discussions. Special thanks go to Jasmin Spekkers, Leonie Diffené, Rebecca Kittel, and Pit Rieger for their invaluable research assistance.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia.

ORCID iD

Franziska Pradel https://orcid.org/0000-0003-4559-0772

Supplemental material

Supplemental material for this article is available online.

Notes

1. For example, 75% of European citizens who follow or participate in debates have experienced hate speech, and about 50% indicate that it makes them hesitate to participate in debates (Special Eurobarometer 452, 2016).
2. See Table S2 in the SI for more details on how the conceptualizations of the study align with practical definitions of hate speech by platforms.
3. About 85% of participants in this study used search engines at least weekly. Please see Figure S2 in the SI for more details.
4. They found that 64% of all their respondents across the countries Germany, Britain, France, Italy, Poland, Spain, and the United States use search engines at least once a day (Dutton et al., 2017).
5. The pre-registration and the pre-analysis plan were published on http://egap.org/registration/6647.
6. Participants were debriefed by describing the aim of the study, and it was stressed again that all scenarios were hypothetical and that statements such as "refugees are criminal" were fabricated and not true. Participants were also provided with contact details for further questions.
7. The susceptibility of some right-leaning individuals to hate speech can depend on the strength of right-wing ideology, as shown here, and may also be influenced by the presence of other

political attitudes, such as populist attitudes. See, for example, SI, Figure S3, where an exploratory analysis considering a mean score ($\alpha = 0.65$) of populist attitudes (Silva et al., 2018) reveals a trend of right-wing populist individuals being more susceptible to hate speech. The SI also further explores heterogeneous treatment effects based on gender, age, education, and Internet skills, uncovering non-significant effects for all analyses.

8. Specifically, Noble (2018) writes,

> By August 2012, Panda (an update to Google's search algorithm) had been released, and pornography was no longer the first series of results for "black girls"; but other girls and women of color, such as Latinas and Asians, were still pornified. (p. 11)

References

Allport GW (1958) *The Nature of Prejudice*. New York: Anchor Books.

Anduiza E and Rico G (2022) Sexism and the far-right vote: the individual dynamics of gender backlash. *American Journal of Political Science*. Epub ahead of print 31 December. DOI: 10.1111/ajps.12759.

Avdagic S and Savage L (2021) Negativity bias: the impact of framing of immigration on welfare state support in Germany, Sweden and the UK. *British Journal of Political Science* 51(2): 624–645.

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

Bar-Ilan J (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): 1581–1589.

Beisch N and Koch W (2022) Aktuelle Aspekte der Internetnutzung in Deutschland ARD/ZDFOnlinestudie: Vier von fünf Personen in Deutschland nutzen täglich das Internet. *Media Perspektiven* 10: 460–470.

Bilewicz M and Soral W (2020) Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology* 41(Suppl. 1): 3–33.

Billig M and Tajfel H (1973) Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology* 3(1): 27–52.

Bogert E, Schecter A and Watson RT (2021) Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports* 11: 8028.

Brader T, Valentino NA and Suhay E (2008) What triggers public opposition to immigration? Anxiety, group cues, and immigration threat. *American Journal of Political Science* 52(4): 959–978.

Brandt MJ and Crawford JT (2020) Worldview conflict and prejudice. *Advances in Experimental Social Psychology* 61: 1–66.

Brehm JW (1966) *A Theory of Psychological Reactance*. Cambridge, MA: Academic Press.

Breyer B (2015) Left-right self-placement (ALLBUS). *ZIS—The Collection Items and Scales for the Social Sciences*. Available at: http://zis.gesis.org/DoiId/zis83

Burgoon M, Alvaro E, Grandpre J, et al. (2002) Revisiting the theory of psychological reactance. Communicating threats to attitudinal freedom. In: Dillard JP and Pfau M (eds) *The Persuasion Handbook*. Thousand Oaks, CA: Sage, pp. 213–232.

Buyse A (2014) Words of violence: fear speech, or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly* 36(4): 779–797.

Chong D and Druckman JN (2007) Framing theory. *Annual Review of Political Science* 10(1): 103–126.

Coppock A and McClellan OA (2019) Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics* 6(1): 1–14.

Costello M, Hawdon J, Bernatzky C, et al. (2019) Social group identity and perceptions of online hate. *Sociological Inquiry* 89(3): 427–452.

Cuddy AJC, Fiske ST, Kwan VSY, et al. (2009) Stereotype content model across cultures: towards universal similarities and some differences. *British Journal of Social Psychology* 48(1): 1–33.

Czymara CS and Dochow S (2018) Mass media and concerns about immigration in Germany in the 21st century: individual-level evidence over 15 years. *European Sociological Review* 34(4): 381–401.

Decker O and Brähler E (2018) *Flucht ins Autoritäre: rechtsextreme Dynamiken in der Mitte der Gesellschaft: die Leipziger Autoritarismus-Studie 2018*. Giessen: Psychosozial-Verlag.

Dennison J and Geddes A (2019) A rising tide? The salience of immigration and the rise of anti-immigration political parties in Western Europe. *The Political Quarterly* 90(1): 107–116.

Deutsche Welle (2017) *AfD, PEGIDA Hold Side-By-Side Events in Dresden*. Bonn: Deutsche Welle.

Dutton WH, Reisdorf B, Dubois E, et al. (2017) Search and politics: the uses and impacts of search in Britain, France, Germany, Italy, Poland, Spain, and the United States. Technical Report, Social Science Research Network. *East Lansing, MI: Quello Center* Working Paper.

Eberl JM, Meltzer CE, Heidenreich T, et al. (2018) The European media discourse on immigration and its effects: a literature review. *Annals of the International Communication Association* 42(3): 207–223.

Fazio LK, Brashier NM, Payne BK, et al. (2015) Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General* 144(5): 993–1002.

Festinger L (1957) *A Theory of Cognitive Dissonance (Reissued by Stanford University Press in 1962, renewed 1985 by author)*. Stanford, CA: Stanford University Press.

Fischer A, Halperin E, Canetti D, et al. (2018) Why we hate. *Emotion Review* 10(4): 309–320.

Fiske ST (1998) Stereotyping, prejudice, and discrimination. In: Gilbert DT, Fiske ST and Lindzey G (eds) *The Handbook of Social Psychology*. Boston, MA: McGraw Hill, pp. 357–411.

Flynn D, Nyhan B and Reifler J (2017) The nature and origins of misperceptions: understanding false and unsupported beliefs about politics: nature and origins of misperceptions. *Political Psychology* 38(Suppl. 1): 127–150.

Forschungsgruppe Wahlen (2020) Most important problem in Germany. Available at: https://www.forschungsgruppe.de/Umfragen/Politbarometer/Langzeitentwicklung-ThemenimUeberblick/PolitikII/#Probl1

Franzmann ST, Giebler H and Poguntke T (2020) It's no longer the economy, stupid! Issue yield at the 2017 German federal election. *West European Politics* 43(3): 610–638.

Gagliardone I, Gal D, Alves T, et al. (2015) *Countering Online Hate Speech*. Bonn: UNESCO Publishing.

Georgiou M and Zaborowski R (2017) *Media Coverage of the "Refugee Crisis": A Cross-European Perspective*. London: Council of Europe.

Gessler T and Hunger S (2022) How the refugee crisis and radical right parties shape party competition on immigration. *Political Science Research and Methods* 10(3): 524–544.

Gidron N, Adams JF and Horne W (2020) *American Affective Polarization in Comparative Perspective (Elements in American Politics)*. Cambridge: Cambridge University Press.

Gilliam FD and Iyengar S (2000) Prime suspects: the influence of local television news on the viewing public. *American Journal of Political Science* 44(3): 560–573.

Gross N (2023) What ChatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences* 12(8): 1–15.

Guess A and Coppock A (2020) Does counter-attitudinal information cause Backlash? Results from three large survey experiments. *British Journal of Political Science* 50(4): 1497–1515.

Gunther AC (1998) The persuasive press inference: effects of mass media on perceived public opinion. *Communication Research* 25(5): 486–504.

Haak F and Schaer P (2022) Auditing search query suggestion bias through recursive algorithm interrogation. *In: 14th ACM web science conference, Barcelona*, 26–29 June, pp. 219–227. New York: ACM.

Haglin K (2017) The limitations of the backfire effect. *Research & Politics* 4(3): 1–5.

Halperin E, Canetti-Nisim D and Hirsch-Hoefler S (2009) The central role of group-based hatred as an emotional antecedent of political intolerance: evidence from Israel. *Political Psychology* 30(1): 93–123.

Hart PS and Nisbet EC (2012) Boomerang effects in science communication: how motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research* 39(6): 701–723.

Hoffman B, Ware J and Shapiro E (2020) Assessing the threat of incel violence. *Studies in Conflict & Terrorism* 43(7): 565–587.

Hsueh M, Yogeeswaran K and Malinen S (2015) "Leave your comment below": can biased online comments influence our own prejudicial attitudes and behaviors? Online comments on prejudice expression. *Human Communication Research* 41(4): 557–576.

Jackob N, Schultz T, Jakobs I, et al. (2023) *Medienvertrauen in Deutschland (Number Band 10951 in Schriftenreihe / Bundeszentrale für politische Bildung)*. Bonn: Bundeszentrale für politische Bildung.

Joachims T, Granka L, Pan B, et al. (2017) Accurately interpreting clickthrough data as implicit feedback. *In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil*, 15–19 August, pp. 154–161. New York: ACM.

Jowell R, Roberts C, Fitzgerald R, et al. (2007) *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*. Thousand Oaks, CA: Sage.

Katz D and Braly KW (1935) Racial prejudice and racial stereotypes. *The Journal of Abnormal and Social Psychology* 30(2): 175–193.

Kay M, Matuszek C and Munson SA (2015) Unequal representation and gender stereotypes in image search results for occupations. *In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, Seoul, Republic of Korea*, 18–23 April, pp. 3819–3828. New York: ACM.

Klaßen A and Geschke D (2019) *Hass im Netz—Wahrnehmung, Beroffenheit und Folgen von Hate Speech im Internet aus Sicht der Thüringer Bevölkerung*. Jena: Institut für Demokratie und Zivilgesellschaft.

Koltsova O, Nikolenko S, Alexeeva S, et al. (2017) Detecting interethnic relations with the data from social media. In: Alexandrov D, Boukhanovsky A, Chugunov A, et al. (eds) *Digital Transformation and Global Society*. Cham: Springer, pp. 16–30.

Kosmidis S and Theocharis Y (2020) Can social media incivility induce enthusiasm? *Public Opinion Quarterly* 84(Suppl. 1): 284–308.

Kunda Z (1990) The case for motivated reasoning. *Psychological Bulletin* 108(3): 480–498.

Lees C (2018) The "alternative for Germany": the rise of right-wing populism at the heart of Europe. *Politics* 38(3): 295–310.

Lewandowsky S, Ecker UKH, Seifert CM, et al. (2012) Misinformation and its correction: continued influence and successful debiasing. *Psychological Science in the Public Interest* 13(3): 106–131.

Liebe U, Meyerhoff J, Kroesen M, et al. (2018) From welcome culture to welcome limits? Uncovering preference changes over time for sheltering refugees in Germany. *PLoS ONE* 13(8): e0199923.

Lindern Jv (2023) Aleph Alpha: Braucht die deutsche Vorzeige-KI mehr Erziehung? *Die Zeit*. Available at: https://www.zeit.de/digital/2023-09/aleph-alpha-luminous-jonas-andrulis-generative-ki-rassismus

Mader M and Schoen H (2019) The European refugee crisis, party competition, and voters' responses in Germany. *West European Politics* 42(1): 67–90.

Mathew B, Dutt R, Goyal P, et al. (2019) Spread of hate speech in online social media. *In: WebSci '19: Proceedings of the 10th ACM conference on web science, Boston, MA*, 30 June–3 July, pp. 173–182. New York: ACM.

McCoy SK and Major B (2003) Group identification moderates emotional responses to perceived prejudice. *Personality and Social Psychology Bulletin* 29(8): 1005–1017.

McIlroy-Young R and Anderson A (2019) From "welcome new gabbers" to the Pittsburgh synagogue shooting: the evolution of gab. *Proceedings of the International AAAI Conference on Web and Social Media* 13: 651–654.

Neubaum G and Krämer NC (2017) Monitoring the opinion of the crowd: psychological mechanisms underlying public opinion perceptions on social media. *Media Psychology* 20(3): 502–531.

Newman B, Merolla JL, Shah S, et al. (2021) The trump effect: an experimental investigation of the emboldening effect of racially inflammatory elite communication. *British Journal of Political Science* 51(3): 1138–1159.

Nikolaev AG, Porpora D, Coffman N, et al. (2023) Hate speech as a form of entertainment: an unexpected support for the gratification hypothesis on Twitter. Atlantic Journal of Communication. Epub ahead of print 30 August. DOI: 10.1080/15456870.2023.2253344.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Noelle-Neumann E (1993) *The Spiral of Silence: Public Opinion, Our Social Skin*. 2nd ed. Chicago, IL: University of Chicago Press.

Nyhan B and Reifler J (2010) When corrections fail: the persistence of political misperceptions. *Political Behavior* 32(2): 303–330.

Oswald ME and Grosjean S (2004) Confirmation bias. In: Pohl RE (ed.) *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. New York: Psychology Press, pp. 79–96.

Otterbacher J, Bates J and Clough P (2017) Competent men and warm women: gender stereotypes and backlash in image search results. In: *Proceedings of the 2017 CHI conference on human factors in computing systems*, Denver, CO, 6–11 May, pp. 6620–6631. New York: ACM.

Otterbacher J, Checco A, Demartini G, et al. (2018) Investigating user perception of gender bias in image search: the role of sexism. In: *SIGIR '18: The 41st international ACM SIGIR conference on research & development in information retrieval, Ann Arbor, MI*, 8–12 July, pp. 933–936. New York: ACM.

Pan B, Hembrooke H, Joachims T, et al. (2007) In Google we trust: users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication* 12(3): 801–823.

Papacharissi Z (2004) Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society* 6(2): 259–283.

Perry B (2001) *In the Name of Hate: Understanding Hate Crimes*. New York: Psychology Press.

Pradel F (2021) Biased representation of politicians in Google and Wikipedia search? The joint effect of party identity, gender identity and elections. *Political Communication* 38(4): 447–478.

Pradel F, Zilinsky J, Kosmidis S, et al. (2024) Toxic speech and limited demand for content moderation on social media. *American Political Science Review*. Epub ahead of print 24 January. DOI: 10.1017/S000305542300134X.

Prinz C and Glöckner-Rist A (2009) ESS Items zu Asylpolitik und Asylbewerbern. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. Available at: http://zis.gesis.org/DoiId/zis143

Proulx T and Major B (2013) A raw deal: heightened liberalism following exposure to anomalous playing cards. *Journal of Social Issues* 69(3): 455–472.

Purcell K, Brenner J and Rainie L (2012) Search engine use 2012. *Pew Research Center*. Available at: https://www.pewresearch.org/internet/2012/03/09/mainfindings-11/

Pyszczynski T, Henthorn C, Motyl M, et al. (2010) Is Obama the anti-Christ? Racial priming, extreme criticisms of Barack Obama, and attitudes toward the 2008 US presidential candidates. *Journal of Experimental Social Psychology* 46(5): 863–866.

Richter T, Kleinschnittger J, Brettfeld K, et al. (2023) *Threat and integration: attitudes towards refugees in Germany*. Technical Report 1. Hamburg: German Institute for Global and Area Studies (GIGA).

Rovenpor DR, Leidner B, Kardos P, et al. (2016) Meaning threat can promote peaceful, not only military-based approaches to intergroup conflict: the moderating role of ingroup glorification: when threat reduces intergroup violence. *European Journal of Social Psychology* 46(5): 544–562.

Schultheiß S and Lewandowski D (2023) Misplaced trust? The relationship between trust, ability to identify commercially influenced results and search engine preference. *Journal of Information Science* 49(3): 609–623.

Schultheiß S, Sünkler S and Lewandowski D (2018) We still trust in Google, but less than 10 years ago: an eye-tracking study. *Information Research: An International Electronic Journal* 23(3): 1–12.

Silva BC, Andreadis I, Anduiza E, et al. (2018) Public opinion surveys: a new scale. In: Hawkins KA, Carlin RE and Littvay L (eds) *The Ideational Approach to Populism*. New York: Routledge, pp. 150–177.

Soroka S and McAdams S (2015) News, politics, and negativity. *Political Communication* 32(1): 1–22.

Special Eurobarometer 452 (2016) *Media Pluralism and Democracy*. Maastricht: European Union.

Steiner M, Magin M, Stark B, et al. (2022) Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. *Information, Communication & Society* 25(2): 217–241.

Sundar SS and Kim J (2019) Machine heuristic: when we trust computers more than humans with our personal information. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, Glasgow, 4–9 May, pp. 1–9. New York: ACM.

Swire-Thompson B, DeGutis J and Lazer D (2020) Searching for the backfire effect: measurement and design considerations. *Journal of Applied Research in Memory and Cognition* 9(3): 286–299.

Tagesspiegel (2023) KI aus Deutschland: Sprachmodell von Aleph Alpha liefert Hitler-Lob und Rassismus. *Tagesspiegel*. Available at: https://interaktiv.tagesspiegel.de/lab/aleph-alphaki-aus-deutschland-biases-vorurteile/

Tajfel H (1970) Experiments in intergroup discrimination. *Scientific American* 223(5): 96–103.

Tajfel H and Turner JC (1979) An integrative theory of intergroup conflict. In: Austin WG and Worchel S (eds) *The Social Psychology of Intergroup Relations*. Monterey, CA: Brooks/Cole, pp. 33–47.

Udupa S and Pohjonen M (2019) Extreme speech and global digital cultures—introduction. *International Journal of Communication* 13: 3049–3067.

Valentino NA (1999) Crime news and the priming of racial attitudes during evaluations of the president. *Public Opinion Quarterly* 63(3): 293–320.

Van Hoof M, Meppelink CS, Moeller J, et al. (2022) Searching differently? How political attitudes impact search queries about political issues. *New Media & Society*. Epub ahead of print 11 July. DOI: 10.1177/14614448221104405.

Waltman M and Haas J (2011) *The Communication of Hate*. Lausanne: Peter Lang.

Waltman MS and Mattheis AA (2017) Understanding hate speech. In: Giles H and Harwood J (eds) *The Oxford Encyclopedia of Intergroup Communication*. Oxford: Oxford University Press.

Wason PC (1968) Reasoning about a rule. *Quarterly Journal of Experimental Psychology* 20(3): 273–281.

Wigger I, Yendell A and Herbert D (2021) The end of "welcome culture"? How the cologne assaults reframed Germany's immigration discourse. *European Journal of Communication* 37(1): 21–47.

Wike R, Stokes B and Simmons K (2016) *Europeans Fear Wave of Refugees Will Mean More Terrorism, Fewer Jobs*. Washington, DC: Pew Research Center.

Wirz DS, Wettstein M, Schulz A, et al. (2018) The effects of right-wing populist communication on emotions and cognitions toward immigrants. *The International Journal of Press/Politics* 23(4): 496–516.

Wood T and Porter E (2019) The elusive backfire effect: mass attitudes' steadfast factual adherence. *Political Behavior* 41(1): 135–163.

Ziegele M, Koehler C and Weber M (2018) Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting & Electronic Media* 62(4): 636–653.

Author biography

Franziska Pradel is a post-doctoral researcher at the Chair of Digital Governance at the Technical University of Munich. Before this position, she worked as a doctoral researcher at the Cologne Center for Comparative Politics at the University of Cologne. She wrote her thesis on "Biased political information in search engines and their effects." Her research interests include political communication, computational social science, experiments, incivility, content moderation, and gender biases on online platforms. Currently, she is working on the effects of incivility and on content moderation preferences.