Table of Contents

[# Summary (300 words -> am Schluss machen) 1](#_Toc157250847)

[# Motivation and background 1](#_Toc157250848)

[### How does my background align with this topic? 2](#_Toc157250849)

[### What skills/knowledge do I want to get through the project, valuable for my future career? 2](#_Toc157250850)

[# Literature Review 2](#_Toc157250851)

[## Education and Political Opinion / Knowledge / Attitudes 2](#_Toc157250852)

[### Educational Attainment, Political Sophistication and Anti-Immigrant Attitudes (Heijden & Verkuyten, 2020) 3](#_Toc157250853)

[## Hate Speech - definition and attitudes 3](#_Toc157250854)

[### Citizen Preferences for Online Hate Speech Regulation (Munzert et al., 2022) 4](#_Toc157250855)

[# Research question 5](#_Toc157250856)

[## Research Question 5](#_Toc157250857)

[## Hypotheses 5](#_Toc157250858)

[# Data and Methods 5](#_Toc157250859)

[## Data 5](#_Toc157250860)

[### Outcome Variables 6](#_Toc157250861)

[### Independent Variable 7](#_Toc157250862)

[### Control Variables 7](#_Toc157250863)

[## Analysis Plan 8](#_Toc157250864)

[### Text Analysis and Classification 8](#_Toc157250865)

[### Statistical Modeling 9](#_Toc157250866)

[### Cross-National and Subgroup Analysis 9](#_Toc157250867)

[### Comment 9](#_Toc157250868)

[### Additional topics of analysis that could be interesting 10](#_Toc157250869)

[# Appendix 10](#_Toc157250870)

# # Summary

Amidst the growing concern over hate speech online and its amplification via social media and generative AI, the following question arises: How inclusive is the discourse on the trade-off between freedom of speech and protection of vulnerable societal groups?

To investigate this within the phenomenon of hate speech on social media, the following research question is addressed: How does educational attainment influence the complexity of individuals' definitions of hate speech and the polarization in their opinion of what should not be allowed to be said on social media?

Research by Heijden & Verkuyten, (2020) on educational attainment and political attitudes suggests that higher educational attainment goes hand in hand with more sophisticated, but also more ideologically defined political views. It is therefore hypothesized that an academic level of education correlates with more complex definitions of hate speech (H1) and that the opinions of academic respondents on speech regulation on social media are more politically polarized than those of non-academics (H2).

To empirically test these hypotheses, the study will analyze data from the “Global Preferences for Hate Speech Moderation” survey, featuring responses from over 19,000 participants across eleven countries. The methodology involves crafting two indicators from the open-text responses in the survey: one for the complexity of individuals’ hate speech definitions and another for the degree of polarization regarding speech regulation preferences.

The analytical approach includes multinomial logistic regression to examine the complexity of hate speech definitions relative to education levels, and linear or ordinal logistic regression to evaluate the polarization of regulatory opinions (depending on the final measurement of polarization).

The analysis will control for variables such as gender, age, political interest, and empathy to ensure a comprehensive understanding of the educational influence. Stratified analyses will break down the data across different national and ideological contexts.

# # Motivation and background

*“Hate speech is one of the most worrying forms of racism and discrimination prevailing across Europe and amplified by the Internet and social media. Hate speech online is the visible tip of the iceberg of intolerance and ethnocentrism.”* (Keen et al., 2020)

Being a social media phenomenon, hate speech is also described as a threat to the functioning of society at large (Solovev & Pröllochs, 2022). The topic of hate speech and questions of its regulation grows even bigger nowadays, seen that generative AI generates content based on already existent online content and is able to reproduce language up to a huge scale. Consequently, a lively discourse has developed around the regulation of hate speech on the internet, also with regard to the implementation of the EU's Digital Services Act.

This work also takes into account that more and more discourses are being discussed in a specific bubble. Has the hate speech discourse reached the wider society? Or is the trade-off between freedom of speech and protection of vulnerable societal groups a purely ivory-tower discourse led by the academic world? Do opposing opinions on the subject exist primarily in higher educational classes?

To answer these questions, it is needed to find out to what extent the formal educational background plays a role in the understanding of hate speech and the opinion of what should not be allowed to be said on the internet.

While the exact relationship between education (especially formal education) for political knowledge and opinion is unclear (see literature review), quite many papers propose to confront the issue of hate speech with more education on the topic (Keen et al., 2020; Estellés and Castellví, 2020; Jubany, Olga, 2016). This could be due to the fact that more cognitive sophistication also makes people more tolerant (Bobo & Licari, 1989).

Contributing to finding out more about the basis of knowledge people have about hate speech and the role that formal education plays in it, this research might help in defining the right institutions and timing for successful intervention.

### ### How does my background align with this topic?

My previous training as a teacher and my former research on civic education has made me aware of the importance of education for political knowledge, awareness, interest, and opinion. During my professional year at the ministry of labor and social affairs, I got in touch with questions of digital governance (e.g., the AI Act, Data Act, DSA) and the importance of social media for public opinion. Based on these experiences, I see a high value in public discourse on online speech regulation.

### ### What skills/knowledge do I want to get through the project, valuable for my future career?

Through this project, I intend to develop expertise in quantitative thesis writing and enhance my statistical proficiency by working with complex survey data. Additionally, I aim to integrate Data Science techniques, such as Natural Language Processing (NLP), into traditional statistical analysis methods, broadening my skill set for future ventures.

# # Literature Review

The context of my research question encompasses questions of educational attainment effects on political knowledge and attitudes, as well as studies on the phenomenon of hate speech, how it is defined and what this definition (or better these definitions) depend(s) on.

Therefore, the relevant papers to look at with more detail can be sorted into two groups: firstly, papers that examine educational effects on political knowledge and opinion, and then papers that look at hate speech definitions and attitudes.

## ## Education and Political Opinion / Knowledge / Attitudes

If knowledge about hate speech is considered political knowledge, existing research would support the assumption that educational attainment does play an important role, given that education is one of the fundamental forces shaping people's political knowledge (Hall, 2018; Weakliem, 2002; Weinschenk & Dawes, 2019). However, existing research is divided as to whether education is an important factor for political opinion (Bobo & Licari, 1989; Chan, 2019; Heijden & Verkuyten, 2020) and political interest or engagement (Highton, 2009; Witschge et al., 2019).

### ### Educational Attainment, Political Sophistication and Anti-Immigrant Attitudes (Heijden & Verkuyten, 2020)

The paper titled "Educational Attainment, Political Sophistication and Anti-Immigrant Attitudes" by Eva van der Heijden and Maykel Verkuyten investigates the relationship between education level, political orientation, and attitudes toward immigrants and refugee policies. The study was conducted in the Netherlands, with a sample of 1,155 Dutch respondents.

The goal of the study was to find out whether two core ideological aspects of political orientation (endorsement of social conformity and acceptance of inequality) are more present within the views of higher than of lower educated individuals, assumed that the former would have a more sophisticated set of political attitudes.

Following this goal, van der Heijden and Verkuyten could show that the impact of political orientation on people’s attitudes towards immigrants and refugee policies differs for higher and lower educated individuals. The authors found a notable association between right-wing political orientation and negative attitudes towards immigrants and restrictive refugee policies, while political orientation is more ideologically defined for individuals with higher education compared to those with lower education.

The study employed a variety of measures including feeling thermometers towards different immigrant groups, Likert scales for attitudes towards refugee policies, and scales for political orientation, social conformity, and acceptance of inequality. The data were analyzed using structural equation modeling.

In relation to my project, this research also establishes connections between educational attainment and political knowledge and opinion. It underscores the significance of ideological sophistication, demonstrating that higher education fosters a more nuanced and coherent political orientation, which subsequently shapes attitudes on socially and politically contentious issues such as immigration and refugee policies. Moreover, it suggests the possibility that higher educational attainment may lead to less tolerant views towards minorities. This insight motivates the inquiry into whether higher education contributes to a more nuanced understanding of hate speech or, conversely, whether it might lead to increased skepticism about its existence, thereby expanding the range of possible interpretations of what constitutes hate speech due to a more polarized perspective on the matter. This research serves as the foundational framework for H1 and H2 in my project.

## ## Hate Speech - definition and attitudes

The most relevant paper in this section is, of course, the previous study on hate speech regulation preferences of Munzert et al. (2022). This work will refer to further papers for more detailed work on hate speech definitions and the phenomenon itself (Izquierdo Montero et al., 2022; Kansok-Dusche et al., 2023; Sellars, 2016; Solovev & Pröllochs, 2022).

The measurement of broadness and narrowness of hate speech definitions will be based on (Kansok-Dusche et al., 2023) and the construction of a polarization score on the work of XY, which is further described in the method section.

### ### Citizen Preferences for Online Hate Speech Regulation (Munzert et al., 2022)

With their vignette study, (Munzert et al., to be published) are analyzing citizen’s preferences for online hate speech regulation (sample size: 2,622 German and U.S. citizens. Respondents were asked to judge artificial but realistic cases of hate speech in terms of offensiveness, hatefulness and actions that should be taken against these posts. The experiment includes a framing and exposure experiment and is embedded in a survey, allowing the authors to examine the experimental findings in the context of basic sociodemographic information like gender and educational background, as well as political views, media usage and questions on empathy, as former research has shown that those characteristics influence people’s attitudes towards hate speech.

A main finding is that type and severity of the messages are most important for people’s evaluations, while contextual factors about the content seem to be less relevant. Respondents generally prefer less extreme measures like deleting hateful messages, but overall there are substantial differences in opinion between gender and ideological subgroups. Tolerance of unpopular opinions is reduced by primer exposure to hateful content.

The study contributes to the debate over content moderation and regulation of online speech – fostering dialogue about the concrete design and implementation of regulatory frameworks like the Digital Services Act (DSA) in the European Union.

The study is divided into three main experiments: Vignette Experiment (Study 1), Framing Experiment (Study 2), and Exposure Experiment (Study 3). The study used data from the Pulse panel, a subset of YouGov's traditional survey panels, to examine perceptions of hate speech. Respondents were selected based on demographic and political targets, with weights estimated using propensity scores. The Vignette Experiment involved constructing unique vignettes to measure perceived offensiveness and hatefulness of messages. The Framing Experiment examined the effects of different frames on respondents' attitudes towards hate speech regulation. The Exposure Experiment combined vignettes with questions to investigate the downstream consequences of hate speech exposure. Statistical analyses, including hierarchical linear modeling and interaction effects, were used to assess the impact of various factors on respondents' preferences and perceptions of hate speech.

In retrospect, “Citizen Preferences for Online Hate Speech Regulation” (Munzert et al., to be published) can be seen as a preliminary study to the new big survey experiment on “Global Preferences for Hate Speech Moderation”, from which the data will be used in this study. The new vignette experiment embedded in a cross-sectional survey expands a very similar approach to more countries. Therefore, findings from the first paper in the U.S. and Germany can be taken as an anker for follow-up research with the new data.

The paper addresses various factors related to perceptions of hate speech. It points to important control variables for my work and comments on existing hate speech definitions. The study highlights that the perceived harm of speech varies from person to person, and knowledge of the concept may influence this perception. It explores the influence of exposure on opinions, finding mixed results in different experiments. The role of educational attainment is examined, with variations observed between the US and Germany, suggesting the importance of exploring this in other regions, like the global south. The study also raises questions about the influence of norms, both in society and educational institutions. Interestingly, untrained respondents appear capable of judging hate speech similarly, prompting inquiries into the necessity of specialized education on the topic. Finally, the authors draw doubts about potential automation of hate speech regulation from their study, which also underscores the importance of examining the role of educational institutions in shaping common views on hate speech and its potential impact.

# # Research question

## ## Research Question

**\*\*How does educational attainment influence the complexity of individuals' definitions of hate speech and the polarization in their opinion of what should not be allowed to be said on social media?\*\***

Possible Subquestions:

* How does this effect differ between countries?
* How does this effect differ for ideological subgroups?

## ## Hypotheses

(H1): People with higher educational attainment (academics) provide more differentiated/complex definitions of hate speech compared to those with lower educational attainment (non-academics).

(H2): Answers from academic respondents on what should not be allowed to be said on social media show greater (political) polarization than those from non-academics.

# # Data and Methods

## ## Data

The data that will be used was obtained for the study “Global Preferences for Hate Speech Moderation” (Munzert et al., to be published), an intervention experiment conducted as a cross-sectional survey in eleven countries (Brazil, Colombia, Germany, India, Indonesia, Nigeria, Philippines, Poland, Turkey, United Kingdom, United States) across selected regions with different cultural context and variation in free speech norms.

Participants were recruited using Facebook Ads and were planned to be balanced across age (18-40 and 41-65+), sex (male, female), and education (below college, at least college education).

(INTEGRATE TABLE Fig in R if this is balances planned)

The dataset contains answers from 19,172 respondents.

Sample Size per Country:

(INTEGRATE TABLE Fig in R)

After a pre-treatment survey, the dataset contains the results of a vignette experiment on hate speech regulation preferences, a framing experiment, and a question order experiment. In my study, I plan to only use survey items that were asked before the experiments and an open text item about what definition people have of hate speech, which was positioned after the experiments. The pre-treatment variables include information on

* **Basic sociodemographic information** (like gender, year of birth, race, being part of a minority)
* **Speech Traits** (perception of individual freedom of speech, opinion if there should be regulation, empathy)
* **Political preferences and behavior** (interest in politics, opinion on free speech, other political opinions and party affinities)
* **Online behavior** (media usage relevant to the issue, how often people share their opinions on the internet, etc.)
* **Speech governance preferences** (Should there be regulation on expressing opinion? Who is responsible?)

The results of the embedded vignette and framing experiments and a question order experiment will not be considered to answer my research question. Thus, they will not be described in detail. Missingness will be dealt with list-wise deletion. In the following subsections, I will describe the outcome variable, independent variable, and control variables in more detail.

### ### Outcome Variables

The following open-text items from the survey will be used as outcome variables:

> „People have different ideas about what constitutes "hate speech." What about you - how would you personally define hate speech?“

> “And, in your own words, what do you think: What - if anything - should not be allowed to say on social media?”

Respondents answered very differently regarding length and content (see also data report and coding scheme). Within the original study, the open-text answers to the question of hate speech definition are already being classified and quantitatively explored regarding common themes in hate speech definitions. The coding scheme for this task contains items on content, sender features/motivation, target scope features, specified target features, and other features of the statement (like, e.g., if the answer provides an example or no definition at all). The coding scheme will be provided as a basis for this work by the scientists carrying out the main study.

To operationalise the open-text answers in such a way that they can be applied to answer the research question, two instrumental variables will be created, measuring differentiation in definitions and political orientation of every statement about what should not be allowed to say on social media (more details in analysis plan).

### ### Independent Variable

The variable measuring educational attainment is available as a 6 to 9-level scale per country. Based on a coarser categorization across countries, an additional educational variable was created within the main study, encoded in three levels:

* Low: did not finish school (yet), or finished school but holds no qualification to pursue education to satisfy university entrance requirements
* Intermediate: finished school with qualification to pursue further education to satisfy university entrance requirements
* High: finished school achieving university entrance requirements, and/or holds university degree and/or post-graduate degree

To analyze differences between respondents with academic versus non-academic backgrounds, a third educational variable will be constructed, combining low and intermediate levels as “non-academic” versus high level as “academic”. To filter out respondents who might still be on the path to acquiring an academic title, participants in the first age group (18-29) will not be considered for the analysis. This would exclude 4,689 respondents, while 14,436 observations with respondents aged 30+ would remain (74 NA’s already excluded).

(still necessary to look at e.g. (Schneider, 2022) for more details on this matter)

### ### Control Variables

Dealing with educational attainment and political knowledge/opinions, the literature suggests controlling for gender (Costello et al., 2019; Cowan & Khatchadourian, 2003; Wilhelm & Joeckel, 2019; Wojatzki et al., 2018), age (Lambe, 2004), political interest (Hall, 2018), empathy (Cowan & Khatchadourian, 2003), social media usage (Celuch et al., 2022; Costello et al., 2019), and experience with HS (Costello et al., 2019; Soral et al., 2018).

Most of these variables were measured as pre-treatment variables:

* Gender (female/male/other)
* Age (year of birth, derived age, collapsed into categories 18-29, 30-49, 50-69, 70+)
* Political interest (measured with four levels and collapsed into low/high)
* Empathy (first principal component of responses to three topic-specific items)
* experience with hate speech
  + experience with hate and offensive speech (score created from the answers on five items)
  + online hostile engagement (score created from the answers on three items)

Social media usage was not measured in the survey, but it can be assumed to correlate strongly with experience with hate speech: the more often people use social media, the more often they can experience the scenarios described.

## ## Analysis Plan

### ### Text Analysis and Classification

To test my hypotheses, I plan to create two constructed indicators that both draw their information from the open text survey item on hate speech definition and serve as outcome variables.

\*\*Instrumental variable “Degree of differentiation/Complexity”\*\*

To measure the basis of knowledge people have about hate speech from their open-text answers on how they define hate speech, several types of answers will be identified manually using different combinations from the items in the coding scheme. These combinations of items on content, sender features/motivation, target scope features, specified target features, and other features of the statement will rely on the framework of Kansok-Dusche (Kansok-Dusche et al., 2023) on broadness and narrowness of hate speech definitions.

Each type will be allocated to a numeric value [0 = (“don’t know”), 1 = (“broad understanding”), 2 = (“middle understanding”), 3 = (“narrow understanding”)], defining the instrumental variable “degree of differentiation/complexity”, whereas a more narrow understanding points to a better knowledge of the public discourse and shows the ability to delimit hate speech from other concepts like bullying (Kansok-Dusche et al., 2023). Items of the coding scheme that are not usually associated with hate speech in academic discourse (e.g., “questions/denies its existence”), will be assigned to the type 0 = “don’t know”.

Using a multi-class classification algorithm (ML), each respondent will be allocated a score on how similar the composition of his hate speech definition is to one of the types (0-3). The type of definition most similar to the respondent’s will determine which of the four groups the respondent is allocated to for further analysis.

ALTERNATIVE/ADDITIONAL APPROACHES:

* Using an unsupervised cluster algorithm to define groups of items in the coding scheme that occur together in the beginning, then identify certain types of answers – advantage could be to find more differentiated types of answers, potentially identifying more dimensions than broadness/narrowness. Would require more methodological research and the approach is riskier in terms of valuable results.
* Using a measure of the complexity of language used in responses, as this can be an indicator of cognitive sophistication.

\*\* Polarization score for “What should not be allowed to be said on social media?”\*\*

To create a score of polarization for every response, the open-text item “open\_allow” on what should not be allowed to be said on social media will be used. Applying sentiment analysis, a sentiment score could be measured, pointing towards a degree of more positive or negative sentiments in statements.

To assess political polarization, I will examine the distribution of sentiment scores across respondents. High polarization would be reflected in a bimodal or multimodal distribution, indicating that respondents are sharply divided in their opinions. Assigning numerical sentiment scores to each response, indicating the intensity of sentiment could help to quantify the degree of polarization within my subgroups.

The score could be measured based on existing research on the topic, e.g.:  
  
\*Measuring Online Political Dialogue: Does Polarization Trigger More Deliberation? (Serrano-Contreras et al., 2020)\*

Serrano-Contreras et al. understand polarization as “a result of differing views on ideological or political issues”, in more detail, they describe that a polarizing statement tends to strongly criticize the opposing opinion, causing the reader not to see the opposing position as legitimate. The authors operationalized the polarization of a comment from the distance between the sentiment analysis of the comments and the median of the sentiment analysis aggregate of all the comments of each text item. Serrano-Contreras et al. recommend a supervised approach for specific topics.

\*#Polar Scores: Measuring partisanship using social media content (Hemphill et al., 2016)\*

As an alternative approach, Hemphill et al. (2016) could give helpful methodological advise. They primary assumption is that a polarized text is highly predictive of the political party of its author. The authors examine how easily it is to predict the political affiliation given the usage of a certain hashtag on X, using a machine learning approach to feature selection. Since the survey data contains information about the party affiliation of people and their position on a left-right scale, a score of predictiveness could be calculated, pointing towards a polarized statement.

To create a score of polarization for every response, the open-text item “open\_allow” on what should not be allowed to be said on social media will be used. A score could be measured based on existing research on the topic, e.g.:  
  
\* Measuring Online Political Dialogue: Does Polarization Trigger More Deliberation? (Serrano-Contreras et al., 2020)  
\* On an Approach to Measure the Level of Polarization of Individuals’ Opinions (Chartishvili et al., 2019)  
\* #Polar Scores: Measuring partisanship using social media content (Hemphill et al., 2016)

### ### Statistical Modeling

I plan to use multinomial logistic regression to model the relationship between educational attainment (academic/non-academic) and the categories of hate speech definition complexity (0-3).

For the polarization score, linear regression or ordinal logistic regression could be suitable, depending on how polarization is quantified at the end.

Control variables such as gender, age, political interest, empathy, and experience with hate speech will be integrated into the models to adjust for potential confounders.

### ### Cross-National and Subgroup Analysis

Conduct stratified analyses to explore variations across countries and ideological subgroups.

Potentially employ interaction terms in regression models to examine differential effects of education on hate speech perception across different demographics.

### ### Comment

If H1 emerges as false: Instead of comparing academics/non-academics, I would suggest comparing two groups where Group A has no or only a broad understanding of the concept (0 or 1), and Group B having a middle or narrow understanding (2 or 3), using the operationalization for H1.

Additional ideas to explore this relationship if it is found to be unfeasible to create a sensemaking polarization score:

1. Analyze the differences in the higher and the lower-educated group (controlled for confounders) in the following items on speech governance preferences
   1. “People should be able to speak their minds freely online/People should be able to feel welcome and safe online”
   2. “Online services should not be responsible for the content users post on their site, even when it’s harassing./ Online services have a responsibility to step in when harassing behavior occurs on their site.”
2. Find users that seem to be “anti-hate-speech”, if their definition contains one of these items of the coding scheme:
   1. Questions/denies existence of hate speech
   2. Questions its importance/relevance
   3. Advocates for free speech
   4. Emphasizes the subjectivity of hate speech   
        
      If several of these items apply, the statement points to a more critical view of the phenomenon itself. Definitions that carry one or more of these properties will be allocated to a critical view on the objective existence of hate speech (“1”), while the others will be assigned a more supportive understanding of the concept (“0”). Then look at the distribution of academics/non-academics in both groups, controlled for confounding variables.

### ### Additional topics of analysis that could be interesting

* Do people who are exposed to online hate know more about Hate Speech? (until now mixed evidence about the impact of exposure, see also (Munzert et al., 2022)
* Do vulnerable groups regarding online hate ((Solovev & Pröllochs, 2022): POC from the Democratic party, White Republicans, and women) know more about Hate Speech?
* Can people who actively produce hate speech define the phenomenon?

# # Appendix

* PAP of the main study
* Survey of the main study (US)
* Coding Scheme