The Spreading of Hostility: Unraveling of Social Norms in Communication

Inauguraldissertation

zur

Erlangung des Doktorgrades

der

Wirtschafts- und Sozialwissenschaftlichen Fakultät

der

Universität zu Köln

2019

presented

by

Amalia Álvarez Benjumea

from

Seville (Spain)

First reviewer: Clemens Kroneberg

Second reviewer: Fabian Winter

Date of oral defense: 22.08.2019

Acknowledgements

This work is the result of my research as a graduate student. Almost four years have passed since I started working on this, and many more since I decided I would love to pursue a career as a sociologist in research. I owe that decision to my years in Barcelona as a bachelor student. There, I learned the joy of discussing sociology as a science to understand human behavior. There was, of course, a lot of late nights and good friends in the mix. Further studies provided me with the tools required to develop this dissertation. I learned a lot about the experimental method at UCL in London, but my ideas about what research in sociology should look like were shaped by the University of Stockholm. The journey continued with the past almost 4 years as a graduate student. I have enjoyed it all the way.

Needless to say, this work is only possible because of the effort and support of other people. My supervisor, Fabian Winter, coauthored two of the three articles in which the chapters in this dissertation are based. He also provided intriguing ideas, and stimulating environment, and support to develop my own ideas. He gave me the opportunity to do this work and treated me as a colleague from the beginning, for which I am really thankful. Working in the "Mechanisms for Normative Change" group has been a fantastic experience and I look forward to continuing working with this team. I also want to thank Clemens Kroneberg for accepting me as a supervisee at the University of Cologne. Finally, thank you to all of those who shared their Ph.D. time with me. Especially those with whom I shared summer schools, beers, and dinners.

I am also very grateful to my family, which has always provided me with infinite support and trust. Especial mention to my sister, who kept visiting me wherever I have been living all these years. She is the person I miss the most. If there is someone I could not have done this without, that is Daniel, my partner, who during these years has patiently woken me up every morning with a cup of coffee. I cannot explain how waking up to that every day feels. He also supported me in all other imaginable ways. We make a great team.

Authorship

The three chapters in this dissertation are based on three separate articles. The first chapter, entitled “Normative change and culture of hate: An experiment in online environments”, has been published together with Fabian Winter with the same title in the European Sociological Review (2018:34(3)). Fabian and I collaborated in the idea, the design of the experiment, the data collection and data analysis process, as well as in the writing. The third chapter, “The Breakdown of Anti-Racist Norms: A Natural Experiment on Normative Uncertainty after Terrorist Attacks” is based in a working paper with the same title also coauthored with Fabian Winter. The article is currently under review. Again, the article is the product of very close collaboration. Fabian and I collaborated in the idea, the design of the experiment, the data collection and data analysis process, as well as in the writing. The second chapter, entitled “Uncovering Hidden Opinions: The Effect of Social Acceptability on Disclosure of Anti-immigrant Views”, is based on a working paper with the same title. The idea, the design of the experiment, the data collection and data analysis process, as well as in the writing was done solely by me as the only author of the article. Nevertheless, to maintain coherence, I use the plural we throughout all the chapters and reserve the singular I for personal opinions in the introduction.

Summary

This work investigates the relationship between social norms, the shared rules that provide the standard of behavior, and online hate speech. We test our hypotheses empirically with three different online field experiments. Each chapter thus addresses a particular perspective of the relation between social norms and hate speech. In the first study, we compare informal verbal sanctions and censoring hateful content as interventions to tackle online hate. The interventions are based on two conceptualizations of social norms commonly found in the literature: i) the observed pattern of behavior or descriptive norm, and ii) informal social sanctions or the injunctive norm. The results suggest that adherence to the social norm in online conversations might be motivated by the observed pattern of behavior. In the second study, we test the assumption that observing an increasing number of norm violations in the local context will result in a decreased willingness to follow the norm, which will eventually result in the breakdown of the norm. In the last study, we explain the rise in online hate speech after terrorist attacks by the terrorist attacks creating a situation of normative uncertainty in which the previous consensus on the social norm against the public expression of hate erodes.

Taken as a whole, the chapters represent an up-to-date general picture of the determinants of how social norms affect hate speech. All the conclusions come from original empirical work. Our data show that highlighting the anti-hate norm results in reduced levels of online hate speech. We also show that the presentation of the norm matters and the observed pattern of behavior is often a powerful normative cue. The descriptive norm seems of key importance for the regulation and they might help to design effective social norm interventions against online hate speech. Different individual characteristics, such as gender, might also affect the way people respond to normative cues. Finally, not only the behavior of others produce normative changes, but events that affect the normative certainty can also impact the anti-hate norm. Particularly events that increase normative uncertainty can amplify influence processes because people resolve the uncertainty by looking at existing patterns of behavior in the context.

Contents

List of Figures

List of Tables

16

Introduction

> *"The reason why many whites do not implement their racial predispositions in all situations is that they have learned to inhibit them in response to the presence of a norm of racial equality"* (Mendelberg, 2001, p. 201)

This work investigates the relationship between social norms and hate speech in online settings. Although hate speech is hardly a new phenomenon, the availability of online social media amplifies it to the point that there is a demand to tackle online hate. Online social media is an opportunity for increased social participation but, at the same time, it opens up to hostility towards vulnerable groups, such as women, the LGBT community, and other minority groups. Online hate speech is now part of the agenda of many European governments.[1] A unique feature of online environments is the lack of formal rules in many online contexts, which emphasizes the role of social norms and makes the study of the relationship between social norms and online hate speech particularly relevant. Furthermore, the relationship between perceived social norms and prejudice expression has been studied almost exclusively in offline settings while similar empirical research in online communities is almost non-existent. Moreover, the nature of online social interaction, in which interactions are long-lived, provides a unique opportunity to study actual behavior instead of relying on self-reported measures.

Social norms are shared rules that provide the standard of behavior. They can be conceptualized as expectations about which behaviors are socially accepted in a specific situation and which are not. (Elster, 1989; Coleman, 1990b; Hechter & Opp, 2001; Bicchieri, 2006). Reactions to norm transgressions range from being frown upon to social sanctioning or even rejection, so individuals are motivated to follow the norms in their social context (Cialdini & Goldstein, 2004; Hechter & Opp, 2001). As with other social behavior, communication is regulated by social norms. Social norms regulate what can or cannot be said in public. Particularly, individuals

[1] The German parliament passed a pioneering 'anti-online hate speech law' in June 2017, which requires social media sites to remove all hate and extremist content: the Net Enforcement Law, *Netzwerkdurchsetzungsgesetz* (Bundesgesetzblatt 2017 Teil I Nr. 61, 07.09.2017, 3352-3355). At a supranational level, the European Commission launched an EU Code of Conduct to prevent and counteract illegal hate speech online. The conduct was released in May 2016 and several online companies, such as Facebook, Twitter or Youtube, joined the initiative as well as European states (see `https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en`, last accessed on 3 April 2019)

might avoid expressing views if they believe they are not approved in their social context. In the case of hate speech, empirical research has repeatedly confirmed the existence of an social norm against its use (e.g., Ivarsflaten, Blinder, & Ford, 2010; Blinder, Ford, & Ivarsflaten, 2013; Crandall, Eshleman, & O'Brien, 2002; Ford, 2008; Blanchard, Crandall, Brigham, & Vaughn, 1994; Mendelberg, 2001; Paluck & Green, 2009).

The idea that the expression of prejudice is subject to normative influence is an old one. As early as in the mid-twentieth century, scholars in social psychology wrote seminal work arguing that the expression of prejudiced attitudes arises from conformity to the normative expectations of a group (Allport, 1979; Sherif & Sherif, 1953). In their work, Sherif and Sherif (1953) describe how the pressure to conform to the group norms could be the cause of the emergence of norms of prejudice expression.[2] A growing body of research in recent years has focused on the role social norms play in shaping the expression of prejudice, particularly how social norms make people avoid its expression in public even when private beliefs remain unaltered (Paluck, 2009a, 2011; Nolan, Schultz, Cialdini, Goldstein, & Griskevicius, 2008), or how social norms shape political discourses (Mendelberg, 2001). Interventions targeting social norms might be more effective in reducing the expression of hate since behavior is generally more closely connected to norms than to personal attitudes (Paluck, 2009b; Tankard & Paluck, 2016).[3]

The public expression of prejudiced, racist or xenophobic views in public became disapproved over the past decades (e.g. Huddy & Feldman, 2009; Dovidio & Gaertner, 1986, 1991; Mendelberg, 2001), particularly since the 1950s (Pettigrew, 1958; Duckitt, 1992; Schuman, Steeh, Bobo, & Krysan, 1997) and the de-legitimization of the biological racism (Rydgren, 2005; Mendelberg, 2001). There was even a decline in open support for political parties with openly xenophobic or anti-LGBT agendas (Ignazi, 1992). An *anti-hate* norm emerged. Views that are in conflict with the anti-hate norm are not tolerated and therefore people refrain to voice them, especially when the norm is brought into focus (e.g., Paluck & Green, 2009; Crandall et al., 2002; Munger, 2016; Álvarez-Benjumea & Winter, 2018). The emergence of the anti-hate norm is also noted very early in research in political science because it generated concerns about the validity of survey measurement of related constructs, such as racial attitudes (Huddy & Feldman, 2009). Survey respondents became less willing to endorse certain beliefs and the "gap between private opinion and public utterance" increased due to the social desirability concerns (Berinsky, 1999).

Several names are used in the literature to refer to the anti-hate norm: egalitarian norm (Huddy & Feldman, 2009; Crandall et al., 2002), norm of racial tolerance (Weber, Lavine, Huddy, & Federico, 2014), anti-racism norm (Ivarsflaten et al., 2010), anti-prejudice norm (Blinder et al., 2013), or tolerant social norm (Paluck & Green, 2009) are among them. Throughout this work, we use the preferred term anti-hate norm because it covers the expression of prejudice towards

[2] Sherif and Sherif (1953) argue that the majority of attitudes are indeed formed in relation to group norms, and maintained by group identification and repeated social interaction. Group norms are therefore the source of appropriate attitudes, and personal attitudes derive from them.

[3] Again, this resonates with Sherif and Sherif (1953) argument. If personal attitudes are "first and foremost group beliefs (social norms)" (Sechrist & Stangor, 2005, p. 168) then changing social norms might be more effective than changing personal attitudes.

different social groups: immigrants and refugees, LGBT, and women. All of them are, at some point, the object of study in this work. We also use the umbrella term *hate speech*, which we define as speech intended to promote hatred on the basis of race, religion, ethnicity, or sexual orientation (Gagliardone, Gal, Alves, & Martinez, 2015). Nevertheless, we also use the terms anti-prejudice norms, and norms of polite or civic expression from time to time. In particular, in the second experiment, the expression anti-racism norm is used because this experiment focuses on the expression of prejudice towards refugees and immigrants.

Yet norms change. The effect of the anti-hate norm is not fixed and variations occur. Often, change happens fast. For example, hearing others publicly express counter-normative views can embolden people to publicize them as well. In experiments, hearing others either condemn or endorse the anti-anti norm (Blanchard et al., 1994; Zitek & Hebl, 2007) led participants to oppose or support these views, as well as, providing consensus information on the anti-hate norm (Stangor, Sechrist, & Jost, 2001). Emphasizing a social norm against the use of hate speech (Cheng, Danescu-Niculescu-Mizil, Leskovec, & Bernstein, 2017) or implementing peer-sanctions (Munger, 2016) reduce the incidence of hate speech in online contexts. Variations can happen in both directions. Generally, highlighting the norm reduces the expression of prejudice (Crandall et al., 2002; Crandall & Stangor, 2005; Shapiro & Neuberg, 2008). On the contrary, presenting a context in which norm transgressions are common produces the opposite effect. The relationship between social norms and the expression of hate has been established by an extensive body of research; however, research on the dynamics of normative change is, until today, virtually non-existent.

In this work, we study different sources of change in the anti-hate norm. Scholars have identified factors contributing to normative change and the topic has remained a core topic in sociology since the beginning of the discipline (Coleman, 1990a, 1990c, 1990b; Ullmann-Margalit, 2015). From the analytical sociology tradition, identifying factors that play a role in normative change generally consists of identifying its *microfoundations.* This means identifying the pathways and mechanisms through which normative change is generated by the actions of the individuals relevant to the situation.[4] This is, in general lines, the followed strategy when developing this work.[5]

As said before, expectations about which behaviors are socially accepted in a specific situation

[4] Identifying the individuals pertinent to an specific situation is of key importance when studying social norms since norms always exist in relation to a group of reference.

[5] The term microfoundations, which refers to the idea that macro phenomena could be traced to micro-level explanation, is largely discussed in the literature of philosophy of the social sciences (see, for example, Little, 1991) and constitutes the base for a mechanism-based explanation, which focuses on the pathways of how phenomena are brought about. The mechanism-based explanation is extensively addressed from sociology (Hedström & Ylikoski, 2010). The idea of a mechanism-based explanation is created in opposition to the covering-law model (Kalter & Kroneberg, 2014), which is seen as impractical by some scholars in sociology (Abbott, 2007). For seminal work, see Hedström, Swedberg, and Hernes (1998) who put together the results of one of the first conferences to address the mechanism based explanation as a foundation for a social explanation, or later works from different authors (Manzo, 2014). Nevertheless, we acknowledge that the concept has much developed since its emergence and that many different accounts of the social mechanism have developed. In this work, we take a weak stance in the idea of causal mechanisms and methodological individualism to include the notion that individuals are socially constructed and locally situated. For a similar account, see Little (2011).

and which are not. When studying the relationship between social norms and hate speech, this works is based on two general ideas that define social norms. First, social appropriateness is embedded in the situation and in the behavior of others relevant to the situation (Bicchieri & McNally, 2018). Second, social norms are interdependent behavior, which means that people prefer to follow a norm conditional on what one expects others to do or to think one should do in such a situation (Bicchieri, 2006, pp. 1-51). It is a fundamental assumption throughout this work that individuals' behaviors are fundamentally interactive and therefore shaped, in part, by the behaviors of others in their social context (Cialdini, Reno, & Kallgren, 1990). This definition of social norms allows us to breakdown the norm in its components and isolate them to better identify sources of normative change.

In the next section, I outline the three research questions this work tries to answer. Each research question corresponds to one chapter of this dissertation. The research questions are described in a general manner in the introduction, but the hypotheses and research questions of each experiment are explained in depth within each chapter.

Research Questions

The first question deals with the effect of different sources of normative information on online hate speech. Previous literature has identified two sources of information: i) what do others normally do, and ii) what happened to those who violated the norm. In the first case, the source of information is a behavioral regularity that can be observed in the context. In the second case, the source is observing sanctions, or the lack thereof, to previous norm violations. Observing the manner in which others have previously reacted to norm violations carries information about what people approve or disapprove in that context. This two distinctive sources of information correspond to what scholar have named descriptive norms, i.e., “what normally happens”, and injunctive norms. i.e., “what others think one ought to do” (e.g., Cialdini et al., 1990; Cialdini & Goldstein, 2004).

Research Question 1: Do different sources of norm communication have different effects on online hate?

Despite this distinction being quite common in the literature, the terms that describe the categories might vary between different authors. For example, Bicchieri (2006) uses the term descriptive norm only in reference to behavioral regularities or conventions. Descriptive norms are based on *empirical expectations*: the belief that a sufficiently large group of people conforms, and will conform, to the convention in a given situation (Bicchieri, 2006, pp. 1-51). In her work, the term injunctive norms is a synonym of social norms and can be used interchangeably. Social norms are those that generate a sense "oughtness".

Nevertheless, almost all authors agree that the line between the two types of norms is blurred as descriptive norms can always become injunctive norms. Other scholars consider that descriptive norms do in fact generate normative expectations in the same manner injunctive norms do (Horne, 2009; Horne, Tinkler, & Przepiorka, 2018; Willer, Kuwabara, & Macy, 2009). I too subscribe to this assumption and therefore see this distinction as just different sources of the norm. Both based on a basic assumption that individuals infer acceptability from the context using the behavior of others as a source of normative information.

The differential effect of injunctive and descriptive norms is tested experimentally using a purpose-built online forum. Building on the distinction between injunctive and descriptive norms, we operationalize the experimental online forum in a way that allows studying whether people learn about norms by observing the prevalent pattern of behavior, i.e., descriptive norm, or by observing peer sanctions to previous norm violations, i.e., injunctive. As in Cialdini et al. (1990) littering experiment[6], I expect a minority of the participants to always use hate speech, another

[6] Cialdini (2009) run a field experiment in which they manipulated situational cues to prime descriptive and injunctive norms about littering in public places. They manipulated the descriptive norm by presenting either clean or littered environments, and test the effect of the environment on the behavior of subsequent participants.

minority to never use it, and a majority of participants that will use it *conditional* on what they observe others do.[7]

The second question builds up in the assumption that the pattern of behavior in the social context, i.e., the descriptive social norm, generates normative expectations and therefore affects the willingness to use hate speech. The idea that a given pattern of behavior affects normative expectations has also been supported by scholars elsewhere (see, for example, Horne, 2009; Horne et al., 2018; Willer et al., 2009), and pointed out as one of the factors contributing to the emergence of a norm (Opp, 2004). This is used as a starting point to investigate how a changing behavioral pattern affects norm compliance, and to attempt to model the social process behind the normative change.

Research Question 2: Does an increasing number of norm violations in the local context lead to the breakdown of the *anti-hate* social norm?

This questions aims to analyze the breakdown of an anti-hate norm in a context in which the number of observed norm violations increases steadily. In this study, I will argue that observing an increasing number of norm violations in the local context will lead individuals to also violate the anti-hate norm, which will eventually result in the breakdown of the norm. The pattern of the norm breakdown will depend on how the individual thresholds for conditionally conform to the new behavioral pattern are distributed in the population, as well as in personal characteristics like motivation to control prejudice (Ratcliff, Lassiter, Markman, & Snyder, 2006). It is a common empirical finding that presenting a context where antisocial behavior is common generates more antisocial behavior, such as littering, stealing, or jaywalking (Cialdini et al., 1990; Keizer, Lindenberg, & Steg, 2008; Keuschnigg & Wolbring, 2015).

The definition of social norms that I have used here includes the conditional compliance assumption, that is, people prefer to follow a norm conditional on what one expects others to do or to think one should do in such a situation. Therefore when a given action is observed to be performed by enough individuals it might generate an expectation that it will be performed, and a feeling of *oughtness* might arise.[8] People have however different thresholds to conform to norms (Sliwka, 2007; Bicchieri, 2006), that is, they have different threshold for what enough individuals actually means. I propose different patterns of normative change based on different assumptions about the distribution of thresholds.

The third question begins with the well-established empirical finding that terrorist attacks are followed by an increase in hate crimes, particularly an increase in online hate speech (e.g., Hanes & Machin, 2014; Awan & Zempi, 2017). Previous literature has attributed the cause of the rise in hate in the aftermath of terrorist attacks to an increase in negative attitudes towards those social groups the perpetrators of the attacks are believed to belong to. Empirical research confirms

[7] See also, *conditional followers* (Bicchieri, 2006, pp. 1-51).

[8] In a similar manner as with *empirical expectations* (Bicchieri, 2006, pp. 1-51)

that terrorist attacks have a profound impact on xenophobic attitudes (Legewie, 2013; Walters, Brown, & Wiedlitzka, 2016; Echebarria-Echabe & Fernández-Guede, 2006; Boomgaarden & de Vreese, 2007). Based on inter-group conflict theory, the perception of the attackers as a threat to the in-group might cause a rise in xenophobic attitudes and prejudice as a response (Riek, Mania, & Gaertner, 2006). The rise in hate speech follows suit, although the link from attitudinal change to hate speech is only assumed and not tested (Hanes & Machin, 2014; Awan & Zempi, 2017).

Research Question 3: Do terrorist attacks increase online hate speech by increasing uncertainty about its normativity?

I will argue that the rise in online hate speech after terrorist attacks is better explained by the effects these attacks have on the normative consensus about the anti-hate norm. Terrorist attacks create a situation of normative uncertainty, which we define as local *anomie.* After the attacks, the previous consensus on the social norm against the public expression of hate erodes and does not provide guidance for behavior. The uncertainty about what behavior is expected makes more receptive to normative cues. As a result, social conformity increases. The resulting effect on hate speech would depend on the available cues in the social context. This *social norm* based mechanism is empirically tested using a combination of a lab-in-the-field and a natural event.

The Methodological Approach: Online Field Experiments

All the studies in this work are different variations of online field experiments. The key feature of experiments is the deliberate manipulation of a particular element of interest to discover the effects it produces on something else (Shadish, Cook, & Campbell, 2002). The aim of experiments is to create a controlled environment in which the links between elements of interest can be observed while holding everything else constant. This setting allows for the identification of cause-effect relations with more clarity than other traditional methods, such as observational data. In observational data the presence of simultaneous events or unmeasured third variables can hamper attempts to establish causal inference, making it difficult to distinguish between causal and spurious correlations (Rauhut & Winter, 2012).

We chose online field experiments over traditional laboratory experiments for several reasons. There are indeed practical reasons like saving time and money. Online experiments typically involve lower participant compensation and less effort from the participants' side (for instance, there is no need for transportation to get to the lab), which allows for recruiting larger samples. Nevertheless, the decision to use online field experiments is based on the opportunity of creating a more realistic environment, recruiting a sufficiently diverse sample, and increasing anonymity of the participants. All these features increase the external validity and the generalizability of the results. Unlike laboratory experiments, online experiments can create a more realistic environment. Online field experiments are "built into the context of the online community under study" (Parigi, Santana, & Cook, 2017, p. 3), and can capture the complexity of the context in which the behavior takes place while still maintaining a controlled environment. This is of particular importance when researching online hate speech and online behavior. This provides better ecological validity and represents a huge advantage over traditional lab experiments.[9]

Recruiting the sample online gave us the opportunity to recruit a more diverse sample than traditional ones (Bader & Keuschnigg, 2018). Traditional sampling for experiments in the social sciences generally involves building up a local subject pool, normally composed of students of similar ages and similar social backgrounds. Furthermore, most of the students tend to come from selected academic subjects, such as economy or psychology. This type of sampling and the inferences from this type of sampling have been previously criticized in the literature (for a critique, see Henrich, Heine, & Norenzayan, 2010). In the case of the online experiments in this work, the recruitment was always crowdsourced via the internet, which ensured a larger sociodemographic variability.

Using the crowdsourced recruiting system means that we never interacted with the participants face to face, further increasing their anonymity. We used the platform Clickworker (`www.clickworker.com`) in all the experiments in this work, a crowdsourcing Internet marketplace. This platform is similar to Amazon MTurk, but with a larger workforce in Germany (our sample was restricted to residents in Germany). Sociodemographic information on the general characteristics of the total population of the workforce of reference is described in Table 1, which was kindly provided by the platform. The increased anonymity, of course, comes with the downside like a decreased control over the experimental setting. Participants are for example, more likely to drop out of the experiment or to pay less attention during the experiment (Keuschnigg, Bader, & Bracher, 2016; J. K. Goodman, Cryder, & Cheema, 2013)

Female	55%
Age	
18-24	28%
25-34	42%
35-44	17%
>45	13%
Employment status	
Student	29%
Employee	26%
Self-employed	15%
Other	20%
N.S	10%

Table 1: Sociodemographic characteristics of the population from where the participants in our experiments were recruited.

[9] Threats to the validity of the experiments will be addressed in the next section.

To sum up, using online field experiments we were able to study the behavior of interest in its natural context and capture the relevant environmental conditions without losing control over the experimental treatments.

The Online Forum

Almost all data used throughout this work were collected in a purpose-built experimental online forum and variations thereof. The design and construction of the online forum thus deserve proper attention. The forum was designed as a platform where invited participants could discuss selected social topics, namely feminism, LGBT, poverty, and refugees. However, only the experiment in chapter 1 uses all topics. The rest of the experiments use different subsets. The forum was designed in three general steps: i) selection of topics and pictures, ii) collection and classification of comments and iii) construction of the different experimental conditions.

The first and second steps are the same for all the experiments and were developed during the pre-experimental stage. The construction of the experiments and the different experimental conditions are idiosyncratic to each experiment and therefore it is explained in depth in each of the chapters. The first two steps common to all experiments are depicted in Figure 1. The online forum is designed to resemble a standard Internet forum.[10] At the beginning of the experiment, they were given a user name and an avatar but remained otherwise anonymous (see Figure 2). Participants were invited to join the conversations and leave comments on topics portrayed in pictures. A screenshot of a typical forum page is shown in Figure 3.

Pictures and topics were selected in a pre-experimental online survey (N=90) from a list of 10 different social topics and 200 pictures.[11] We chose topics and pictures identified as "potentially controversial for discussion" by the respondents in the pre-experimental survey to ensure that all topics were, to some extent, subject to public debate.[12] In total, nine pictures illustrating four topics were selected: three pictures on feminism, two pictures on LGBT rights, three pictures on refugees and multiculturalism, and one picture representing poverty. In a pre-experimental session, we made our forum available online and collected comments on the pictures. We collected a total of 910 comments. All pre-experimental sessions were conducted using workers recruited from Clickworker.

[10] The forum was created using Otree (Chen, Schonger, & Wickens, 2016), a software for economics experiments.

[11] The images were obtained from Twitter and Google during March 2016 and we used a set of tags and keywords to collect them. Both German and English were used in the search, as both languages are often used on German social media. The following terms and derivatives were used: Sharia, Multiculturalism, Terrorism, Transgender, Gay, Sexism, Discrimination, Refugees, *Aufschrei*, Immigration, Homosexuality, *Einwanderung*, Diversity, Queer, Begging, Atheism, Islamization, Religion, *Tolerist* from different online platforms

[12] The survey was conducted in April 2016.

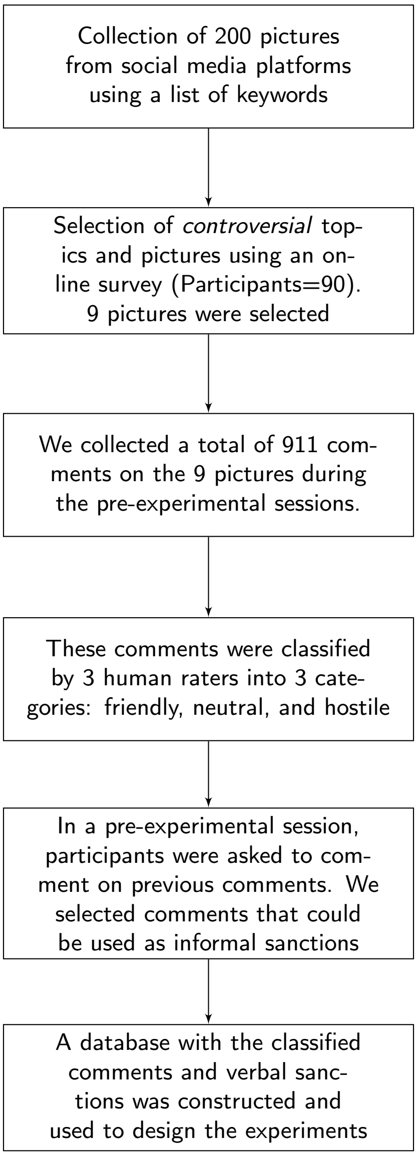
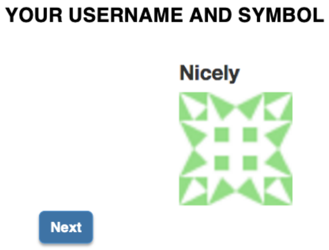

Collection of 200 pictures from social media platforms using a list of keywords
Selection of *controversial* topics and pictures using an online survey (Participants=90). 9 pictures were selected
We collected a total of 911 comments on the 9 pictures during the pre-experimental sessions.
These comments were classified by 3 human raters into 3 categories: friendly, neutral, and hostile
In a pre-experimental session, participants were asked to comment on previous comments. We selected comments that could be used as informal sanctions
A database with the classified comments and verbal sanctions was constructed and used to design the experiments

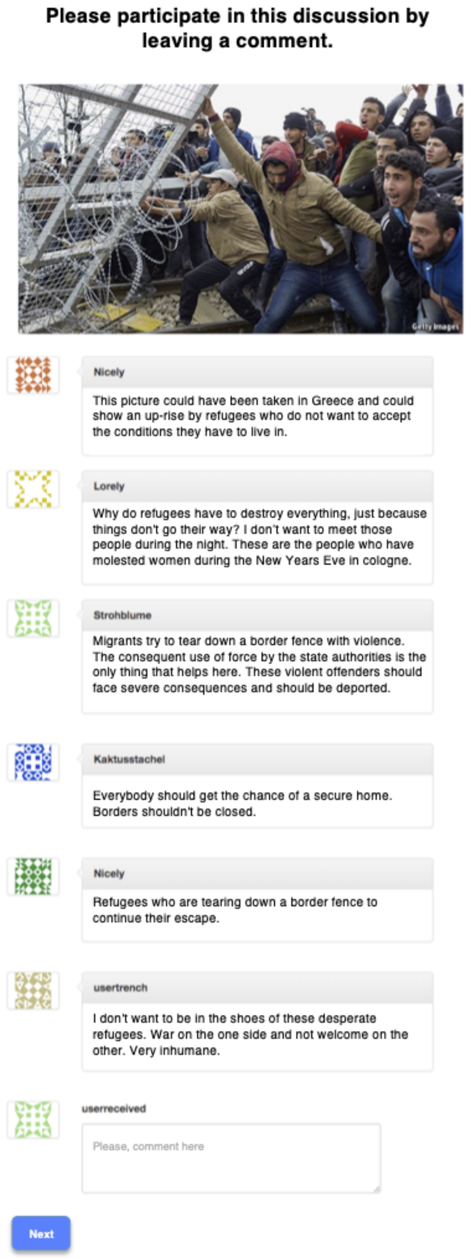
Figure 1: Flowchart of the design and preparation of the experimental online forum during the pre-experimental phase.


YOUR USERNAME AND SYMBOL

Nicely

Figure 2: Screenshot of the assignment of username and avatar to the participants at the beginning of the experiment (own translation. In German in the original.)

The comments collected in this first pre-experimental session were rated by three independent raters using a *hate* score (1 to 9 scale), and classified into 3 categories: friendly (<2), neutral (4.5-5.5), and hostile (>7.5). In a second pre-experimental session using workers, we collected replies



Please participate in this discussion by leaving a comment.

Nicely

This picture could have been taken in Greece and could show an up-rise by refugees who do not want to accept the conditions they have to live in.

Lorely

Why do refugees have to destroy everything, just because things don't go their way? I don't want to meet those people during the night. These are the people who have molested women during the New Years Eve in cologne.

Strohblume

Migrants try to tear down a border fence with violence. The consequent use of force by the state authorities is the only thing that helps here. These violent offenders should face severe consequences and should be deported.

Kaktusstachel

Everybody should get the chance of a secure home. Borders shouldn't be closed.

Nicely

Refugees who are tearing down a border fence to continue their escape.

usertrench

I don't want to be in the shoes of these desperate refugees. War on the one side and not welcome on the other. Very inhumane.

userreceived

Figure 3: Screenshot of a prototype experimental condition (own translation. In German in the original.)

to hateful comments, which could be used as verbal sanctions to construct one experimental condition in the experiment in chapter 1.

Raters were provided with each comment and the following question:

> **Is the comment friendly or hostile towards the group represented in the picture?**
> *(Give a number from 1 to 9 where 1 means very friendly and 9 means very hostile)* .

Comments with a score of 1 are very friendly in language and express a positive opinion. The scores were averaged to construct a *mean hate score* and used to classify the comments. This system applies to the comments collected for the design of the experimental forum. The analysis of the comments collected in the experiments and used as the dependent variable might vary from experiment to experiment and therefore is explained in each chapter. Nevertheless, because of the nature of hate speech, its subtleties and context-dependency, we always used human raters to classify the comments. This strategy was deemed better than a more automatic manner like sentiment analysis, and more comprehensive than dictionary-based classifications.

Validity of the Experimental Design

Experiments have been increasingly used in sociology in the past years. Nevertheless, experiments still are met with skepticism about their validity. In this section, I address possible threats to the validity of our experiments and explain how we tackle them.

Internal Validity Throughout all the steps in the design of the experimental forum, we strive for a high degree of internal validity. In contrast to previous research, the experimental approach allows us to study the production of hate speech under controlled conditions. Previous studies have mainly used observational data, such as data collected from social platforms like Twitter or Facebook. Observational data present several sources of variation that make it difficult to disentangle different competing mechanisms. This would potentially jeopardize the proper identification of the treatment effect. In our forum, we created conditions in which only one feature was manipulated, isolating the treatment effect from other factors that could have played a role. For example, users in those communities see content that has already been filtered because they often self-selected themselves into contexts. We randomly assigned participants to the conditions to study their effects and avoid self-selection effects.

Our design also avoids the high path dependency and endogeneity in observational data when studying influence processes. Because of the circular nature of peer influence, the generation of observational data is particularly is particularly prone to high endogeneity (Angrist, 2014). For instance, in a natural online forum where people respond to each other and history matters a lot. In our forum, participants could not see what other participants immediately before them had commented, but only the comments we had previously selected to create the different conditions. This ensured that all individual observations collected were independent. All of these features increased the internal validity of the experiments in this book and the feasibility of inference of the treatment effects.

External Validity As discussed previously, online experimentation introduces further advantages, such as increased anonymity and a recreation of the natural context where the behavior of interest normally occurs, which increases ecological and external validity (Shadish et al., 2002; Rauhut & Winter, 2012). Nevertheless, there could still be some skepticism, particularly concerning the participants in our forum and their motivation to participate in the forum. An exacerbated social desirability concern might affect the estimation of the magnitude of the treatment effect. We took several measures in order to address this limitation: i) participants remained anonymous and no personal data were collected, ii) participation was always voluntary, and iii) payment did not depend on performance. We believe that these design choices create enough detachment between the forum and the marketplace, and the forum and the experimenters. To avoid further demand effects, the exact purpose of the experiment was unbeknownst to the participants.[13]

Concerning generalizability of the results, the use of online convenience samples could also raise concerns. However, it has been shown that online panels are problematic only to the extent that treatment effects differ between the online sample and the population of interest (Coppock, Leeper, & Mullinix, 2018). Since we do not expect our online sample to perceive the environments any differently than the general population, we believe there is a sufficient overlap to allow an interpretation of the results of the field experiment to people that normally participate in online discussions.

Overview of the Chapters

In **Chapter 1**, the online experiment investigates the impact of the perceived social norm on online hate speech, and measure the causal effect of specific social norm interventions: counter-speaking (informal verbal sanctions) and censoring (deleting hateful content). The interventions are based on the assumption that individuals infer acceptability from the context using previous actions of others in the context as a source of normative information. The interventions are based on two conceptualizations of social norms commonly found in the literature: 1) what others normally do, i.e., descriptive norms; and 2) what happened to those who violated the norm, i.e., injunctive norms.

The experimental treatments are also similar to interventions found in real online settings, which try to emphasize a norm against the use of hate speech by using either a community-driven approach or censoring hate content. The former relies on the feedback produced by peers to reduce hate speech through informal peer punishment. The latter implies directly deleting hate content. To test these two different approaches we designed an online randomized experiment that resemble an online social forum where participants are asked to engage in discussion about pictures of selected social topics. We assigned participants to three different treatments aimed to manipulate participants perception of an anti-hate speech norm. In the baseline treatment, participants could see a mix of comments with different levels of hate speech. In the censored treatment, we delete the hate comments and present a censored environment. In the counter-speaking treatment, we present a forum where the hate comments are sanctioned by the community. We compare the resulting level of hate speech in different conditions. We thus compare the effectiveness of different normative interventions aimed at reducing hate speech.

Participants were significantly less likely to use hate speech when prior hate content had been moderately censored. Our results suggest that compliance with the social norm in online conversations might, in fact, be motivated by descriptive norms rather than injunctive norms. In this chapter, we present some of the first experimental evidence investigating the social determinants of hate speech in online communities. The results could advance the understanding of the

[13] Furthermore, other sampling strategies, such as university students would have not solved this problem. They could also lead to an increased social desirability bias and increase the likelihood of demand effects due to the physical presence of the experimenters.

micro-mechanisms that regulate hate speech. Also, such findings can guide future interventions in online communities that help prevent the spread of hate.

In **Chapter 2**, we take descriptive norms, the idea that the predominant pattern of behavior in the social context serves as the normative reference, as a starting point. We examine how observing increasing norm transgressions in the local context prompts individuals to also violate the anti-hate norm. The experiment focuses on the dynamics of the normative change, and test two different hypotheses of how the dynamics of change could be. The first hypothesis states that norm violations will increase steadily as the number of observed norm-transgressions increase, the linear relation hypothesis. The second hypothesis is based on the idea that people will only violate the norm once a certain threshold has been met, the threshold hypothesis. We also test whether the extent of normative influence depends on the gender of the individual (hypothesis 3).

The results in the experiment in chapter 1 suggest that adherence to the anti-hate social norm might be be motivated by the descriptive norm. This means individuals infer acceptability from the context using previous actions as normative cues. This chapter builds on this idea and explores the dynamics of how the anti-hate norm unravels: how people react to norm violations, how many norm violations break the norm, and the moderating effect of different norm sensitivities. In doing so, we designed an online experiment (N=2283) in which participants were invited to an online forum to discuss immigration issues. We manipulate the social acceptability of expressing prejudice by increasing the proportion of comments considered violations of the anti-hate norm in each consecutive forum page in an exposed experimental condition, whereas in the unexposed condition the tone of the comments remains fixed. A behavioral measure is used to asses normative change. We ask participants to donate to either an anti-immigrant or pro-immigrant organization: Alternative for Germany (AfD), or Pro Asyl. Participants are given the possibility to make a donation of 1 euro. Agreeing to the donation does not entail any monetary cost and therefore it only carries a signaling value.

The treatment conditions vary along three dimensions: i) the type of organization, ii) the number of comments each participant observes in the online forum before the decision, and iii) the fraction of those comments that are hateful. Across systematic variations thereof, we measure how the proportion of norm violations of the anti-hate norm influences the decision to donate. To test the effect of anti-immigrant comments in the decision to donate, the participants were randomized among five different donation decision points at different stages.

The experimental design in this chapter resembles a dose-response model. Besides a comparison to the baseline, exposed and unexposed groups are also analyzed. Furthermore, in this experiment, I introduced a *nonmanipulative* trait, i.e., gender, as a moderator of the treatment effects. Overall, participants are more willing to donate to the pro-immigration organization than to the anti-immigration. Women are particularly reluctant to donate to the anti-immigrant option and reduced even more their donations when the anti-immigrant comments raised normative concerns. This result resonates with the established empirical finding that women, in general, show less support for parties with openly racist political agendas (Harteveld & Ivarsflaten, 2018). Results in this paper can help explain how changes in the normativity of openly expressing xenophobic views can impact the success of right-wing populist parties, and how the anti-hate norm could potentially prevent them to gain further support.

In **Chapter 3**, we examine the dynamics of online hate speech before and after two terrorist attacks in Germany in our experimental online forum. Terrorist attacks can have profound consequences for the erosion of social norms, yet the causes of this erosion are not well understood. We argue that these attacks create substantial uncertainty about whether the anti-hate norm still holds. Observing breaches of these norms then leads people to more readily express their own anti-immigrant attitudes as compared to a context where these norms are not challenged. To test our theory, we examine (i) the impact of terrorist attacks on the level of hate speech against refugees in online discussions, and (ii) how the effect of terrorist attacks depends on the uncertainty of social norms of prejudice expression.

We exploit a natural experiment setting, the occurrence of two consecutive terrorist attacks in Germany, to offer an estimation of the effect of the terror attacks on online hate speech in our forum. This experiment is, therefore, a combination of a natural event, i.e., the terrorist attacks, and the lab-in-the-field experiment. We compare our proposed *social norm* mechanism to the *attitudinal change* mechanism commonly found in previous research. The experiment compares the effect of the terrorist attacks in contexts where a descriptive norm against the use of hate speech is emphasized, i.e. participants observe only neutral or positive comments towards a minority group, to contexts in which the norm is ambiguous because participants observe also anti-minority comments. Hateful comments towards refugees in the experimental online forum, but not towards other minority groups (i.e., gender rights), increased as a result of the attacks. Observing anti-immigrant comments had a considerable impact on our participant's own comments after the attacks while observing anti-gender-rights comments had not.

Achievements in a Nutshell

Each chapter of this work addresses a particular perspective of the relation between social norms and hate speech. Although all chapters report the results of online field experiments, each experiment also has several variations in the experimental design. In chapter 1, the experiment tests the effect of different types of norms, i.e., injunctive and descriptive norms. The experiment is also a test for different interventions to tackle online hate speech. In chapter 2, the experiment uses a behavioral measure, i.e., donations, to study how the anti-hate norm might breakdown when a pattern of norm-violations is presented. In chapter 3, a combination of a lab-in-the-field and a natural event is studied to show how terrorist attacks can impact online hate via an elevated uncertainty about the pertinence of the social norm. The chapters, taken as a whole, represent an up-to-date general picture of the determinants of how social norms affect the hate speech. In each chapter, the conclusion from the findings is elaborated in depth. Nevertheless, several general conclusions of the thesis as a whole can be described.

The work presented here demonstrates that the expression of prejudice, and online hate speech in particular, are affected by relevant social norms. In environments where an anti-hate norm was highlighted, hate speech was reduced. The anti-hate norm affected the use of hate speech even when individual attitudes remained unchanged. This demonstrates a direct relationship between the perceived social norm and the use of hate speech, and opens the way to develop normative interventions to tackle online hate when needed. Furthermore, the descriptive norm, i.e., the observed pattern of behavior, seems of key importance for the regulation of online hate speech. In situations where different normative sources are in conflict, descriptive norms might prevail. Descriptive norms might help to design effective social norm interventions against online hate speech because they effectively motivate norm compliance with the anti-hate norm.

We also find that different individual characteristics might affect the way people respond to normative cues. In this work, gender has been the only moderator of the effect tested. Men and women respond differently to the transgressions of the anti-hate norm. Women show in general more reluctance than men to support anti-immigration discourses, but they also respond to observing norm violations by displaying behavior consistent with the anti-hate norm. In this case, gender might be a proxy for motivation to control prejudice. Further research in individual determinants of reactions to normative cues should be addressed by future research.

Finally, not only the behavior of others produce normative changes. Some type of disruptive events can also impact the anti-hate norm. In this work, we identified a mechanism whereby terrorist attacks might increase uncertainty about what is socially accepted, which eventually produces changes in the overall level of hate speech: terrorist attacks increase the uncertainty about the norm, i.e., the local anomie, which amplifies influence dynamics. This means the reactions depend on the existing cues. A context that normalized hate after terrorist attacks might lead to a cascade of hate speech, and eventually the breakdown of the norm. A context where hate content is limited will not have this effect. The magnifying effect of uncertainty in the influence of descriptive norms might also mean that descriptive norms can provide cues for injunctive norms and, therefore, the distinction between injunctive and descriptive norms is not as sharp as it has been generally assumed.

There are of course many open questions. Future work is needed to identify further factors contributing to normative change or how other conditions, such as social identity or group identification, might affect the adoption and spreading of the anti-hate norm. Furthermore, future research should address whether changes in the anti-hate norm are either long-lasting or short-lived, which factors influence the persistence of these changes, and whether changes in both directions are symmetrical. Are the dynamics of normative breakdown similar to the dynamics of building the norm up again? Many questions remain open. I believe, however, this work opens up new directions for future research since it was written with the hope it would inspire others to see this topic as the important, exciting endeavor it has been to me.

Chapter 1

Normative Change and Culture of Hate: An Experiment in Online Environments[1]

The rise of online social interaction has opened the way for increased social participation. At the same time it has unlocked new gates to express hostility, making engagement harder for vulnerable groups, such as women, the LGBT community[2], or other minority groups (Kennedy & Taylor, 2010; Mantilla, 2013). This behavior is commonly referred to as hate speech. Hate speech is defined as speech intended to promote hatred on the basis of race, religion, ethnicity, or sexual orientation. It is closely related to other types of online antisocial behavior, such as online harassment and trolling (Binns, 2012), since people who engage in these types of behavior often make use of such methods. In this article, however, we will limit the analyses to hate speech, as we understand trolling as an umbrella term for different antisocial behaviors. We define hate speech as hostile behavior and "antagonism towards people" (Gagliardone et al., 2015, p. 11) who are part of a stigmatized social group. The concept is, therefore, close to prejudice expression.[3]

Hate speech may cause fear (Hinduja & Patchin, 2007) and push people into withdrawing from the public debate, therefore harming free speech (Henson, Reyns, & Fisher, 2013) and contributing to a toxic online environment. Social platforms and organizations established to combat hate speech have recognized that online hateful content is increasingly common.[4] As a result, many governments and online media platforms have implemented diverse campaigns and inter-

[1] This chapter has been published together with Fabian Winter with the same title in the *European Sociological Review* (2018:34(3))

[2] Lesbian, Gay, Bisexual and Transgender.

[3] We will use the terms hate, hostility, and prejudice expression interchangeably in the text.

[4] UN Human Rights Council Special Rapporteur on Minority Issues (HRC, 2015) or Council of Europe, Mapping study on projects against hate speech online (15 April 2012). For some statistics, see Hate Base (`http://hatebase.org`).

ventions to tackle online hate speech.[5] Efforts against online hate speech often involve favoring counter-speaking (flagging, reporting, etc.) or censoring the hate content (Citron & Norton, 2011). The theoretical and policy-making importance of these interventions has not yet been well understood.

We conducted a novel experiment to further our understanding of the underlying mechanism of hate speech. We tested whether decreasing social acceptability of hostile comments in a forum could prevent hate expression, and measured the causal effect of specific interventions. We used interventions designed to reduce hate speech in online environments: censoring hate content and counter speech. [6] We designed an online forum and invited participants to join and engage in conversation on current social topics. We chose an online forum because online discussions are the basis of many social platforms on the Internet. Our experiment manipulates the comments participants could see before giving their own comments. The censoring treatment is a top-down approach that consists of censoring hate content and presenting an environment where prior hate comments are not observed. In the counter-speaking condition, the hate comments are presented with comments calling attention to the unacceptability of hate speech. The experiment was conducted with 180 subjects recruited from a crowdsourcing platform. We collected the comments from conversations in the forum and compared the level of hostility of the comments and instances of hate resulting across the conditions.

The interventions are based on the theoretical claim that social acceptability can be inferred from previous action. This claim is based on the observation that presenting a context where antisocial behavior is common brings about more antisocial behavior, such as littering, stealing, or jaywalking (Cialdini et al., 1990; Keizer et al., 2008; Keuschnigg & Wolbring, 2015). A similar process has been found in online contexts, where prior troll comments affect the likelihood of subsequent trolling (Cheng et al., 2017). This cascading dynamic is linked to a process of spreading norm violations: people learn from each other which kind of behavior is approved and which behavior people are to expect in particular situations.

When people observe that others have violated a certain social norm, such as expressing hate-

[5] Concerns about hate speech and violence can be linked to responses at various levels. Digital platforms, for instance, allow for different responses. In many cases, platforms present some type of moderation process (E. Goodman & Cherubini, 2013). Community guidelines, such as in Facebook (`https://www.facebook.com/communitystandards`) and Youtube (`https://www.youtube.com/yt/policyandsafety/communityguidelines.html`), are also common. International initiatives to keep track of hate speech across networks have also emerged, such as HateBase and Fight Against Hate (Gagliardone et al., 2015). At the national level, countries like Germany have made huge efforts to combat online hate speech. On June 2017, Germany approved a law, the Netzwerkdurchsetzungsgesetz, which requires social media sites to remove all hate and extremist content (Bundesgesetzblatt 2017 Teil I Nr. 61, 07.09.2017, 3352-3355).

[6] Different interventions have been discussed in the literature. E. Goodman and Cherubini (2013), for example, refer to pre and post-moderation of content as a strategy for creating better conversations online. Kraut et al. (2012) discuss evidence-based recommendations to design better online platforms. Among the strategies presented, using descriptive norms to tackle online hate speech is discussed (Kraut et al., 2012, p. 13). Furthermore, Schieb and Preuss (2016) present empirical evidence on counter speech as a measure for tackling hate speech on online platforms such as Facebook. We use this discussion as motivation for our treatments and construct them as ideal types of existing interventions, which help us identify clean treatment effects. For evidence-based general recommendations on how to design online communities, please see Kraut et al. (2012) and E. Goodman and Cherubini (2013).

ful views, they are more likely to transgress it because they perceive the behavior as socially accepted. The opposite should hold true: reducing the social acceptability of hateful behavior online might reduce the willingness of individuals to engage in hate speech. This relationship between perceived social norms and prejudice expression in offline settings has been widely studied (e.g., Pettigrew, 1991; Paluck & Green, 2009). Highlighting a majority norm, or a perceived consensus against the expression of prejudice, reduces people's willingness to express prejudice (Crandall et al., 2002; Crandall & Stangor, 2005; Stangor et al., 2001; Shapiro & Neuberg, 2008). Parallel empirical research in online communities is still scarce, with some evidence of the effect of perceived norms on hate expression, such as the effect of promoting a norm through social sanction (Munger, 2016) or reminding participants of etiquette rules (Matias, 2016).

The experimental approach allows us to study the production of hate speech under very controlled conditions. Data from observational studies might present several sources of variation that make it difficult to disentangle different competing mechanisms. While users in those communities have already filtered content and self-selected themselves into contexts, we created and randomly assigned participants to those conditions to study their effects. Online experimentation introduces further advantages, such as increased anonymity – both among participants and towards the experimenter – and a recreation of the natural context where the behavior of interest normally occurs, which increases ecological and external validity (Shadish et al., 2002; Rauhut & Winter, 2012).

1.1 Theory and Hypotheses

Social norms are shared rules that provide the standard of behavior within a wide range of settings (Elster, 1989; Coleman, 1990a, 1990c; Hechter & Opp, 2001; Bicchieri, 2006), and the behavior that people are to expect in particular situations.[7] Individuals are motivated to understand norms in their social context because they care about how they are perceived by others (DellaVigna, List, Malmendier, & Rao, 2016), to avoid rejection (Cialdini & Goldstein, 2004), or to avoid further sanctioning (Hechter & Opp, 2001). Norms are usually not clearly determined and individuals rely on their subjective perceptions of norms (Tankard & Paluck, 2016); people use the behavior of others as a source of information about social norms, and follow a norm conditional on their expectations about how others behave and how others believe one should behave in similar situations (Bicchieri, 2006).

The way we communicate is also regulated by social norms; in particular, individuals avoid publicly expressing views if they believe they are not popular in their social context (Bursztyn, Egorov, & Fiorin, 2017; Cialdini & Goldstein, 2004; Cialdini & Trost, 1998). Likewise, prejudice is subject to similar normative influence.[8] There is evidence of a norm against public expression

[7] Social norms might take the form of quick quasi-automatic answers or completely developed actions, since they are often grounded in "scripted sequences of behavior" (Bicchieri & McNally, 2018, p. 2).

[8] Allport (1979) and Sherif and Sherif (1953) wrote seminal texts arguing that the majority of prejudiced attitudes arose from conformity to social normative expectations. In their work, Sherif and Sherif (1953) describe

of hate in Europe (Ivarsflaten et al., 2010; Blinder et al., 2013), which makes an expression of prejudice more likely in a private than in a public context (Ford, 2008).[9]

Because norms are interdependent, information about the behavior of others is pivotal for normative change. For example, providing consensus information over negative stereotypes (Stangor et al., 2001) can reduce the adherence of people to prejudiced views. Events such as elections can have an effect, as they disclose information on the prevalence of certain opinions and induce changes in the perception of social acceptability. Bursztyn et al. (2017) argue that the 2016 election results in the USA causally increased individuals' perception of the social acceptability of anti-immigration views and "their willingness to publicly express them".

Observing the behavior of others around us is a key source of information on established social norms (Bicchieri, 2006). Lab experiments show that prejudice expression can be reduced by manipulating "normative acceptability of prejudice" (Blanchard et al., 1994, p. 362), showing consensus information over negative stereotypes (Stangor et al., 2001), or hearing others endorse an anti-prejudice norm (Zitek & Hebl, 2007). Further experiments found that individuals not only suppress prejudice expression, but also are more likely to oppose discrimination immediately after hearing someone else do so (Cialdini & Trost, 1998).

1.1.1 Conveying Information about Appropriateness

We have argued that individuals infer acceptability from the context using the behavior of other actors within it as a source of normative information. Previous literature has identified two sources of information: 1) what do others normally do, and 2) what happened to those who violated the norm. This is the distinction between "what normally happens", i.e., descriptive norms, and "what others think one ought to do", i.e., injunctive norms (e.g., Cialdini et al., 1990; Cialdini & Goldstein, 2004; Bicchieri & Xiao, 2009). Descriptive norms act as coordination devices of "normal behavior" (e.g., Bicchieri, 2006; Krupka & Weber, 2013), whereas injunctive norms act as an "oughtness rule" (e.g., Hechter & Opp, 2001). Situational triggering cues can also increase the saliency of normative information and reduce ambiguity about the appropriateness of a certain type of behavior. Actions that stand out, such as observing someone punishing, draw attention to the existing norm. The implications for online behavior are straightforward. While observing prevalent online behavior illustrates the descriptive norm, observing responses to those behaviors teaches the injunctive norms.

Building on the distinction between injunctive and descriptive norms, we operationalize the online setting in a way that allows us to study whether people learn about norms by observing them (descriptive norm mechanism) or by observing norm violations being sanctioned (injunctive

the development of prejudice-expression norms as the result of the pressure that the group places on individuals to conform to the group norms.

[9] Rejection of public expression of prejudice has been generally increasing in the last decades in many western societies (Pettigrew, 1958; Duckitt, 1992), although differences between countries are broad. For example, European countries tend to have a stricter view of what can be considered hate speech, whereas in North America more weight is given to free speech (Pettigrew, 1958; Duckitt, 1992; Dovidio & Gaertner, 1986)

norm mechanism). To do so, we adapt interventions designed to reduce hate speech in online environments: censoring hate content and letting peers verbally sanction it. Censoring hostile content biases the individual's perception of the prevalence of hate speech, i.e., the descriptive norm. If norms are followed because individuals perceive that a majority adheres to it, expectations are that people will not make use of hate speech. This mechanism predicts a positive relationship between the subject's action and what she observes others have done, thus:

Hypothesis 1a *Removing examples of hate speech in the online context, therefore decreasing its observed prevalence, will accentuate a descriptive norm and lead to less hostile content.*

However, people making the choices are also heterogeneous, i.e., they require different amounts of social pressure to elicit a particular response. It is possible that merely deleting hate speech instances would not be a strong enough signal of an anti-hate norm for a majority of individuals. Thus, we have that:

Hypothesis 1b *Presenting only cases of friendly speech, therefore increasing its observing prevalence, will accentuate a descriptive norm and lead to less hostile content and fewer instances of hate.*

Observing explicit counter-comments to hate content, i.e., verbal sanctions to a hateful comment, signals injunctive norms and clarifies the behavior that is believed to be appropriate. We decided to use verbal sanctions because they fit naturally into an online conversation setting. Also, verbal sanctions, such those in online firestorms, are used as online normative enforcement (Rost, Stahel, & Frey, 2016), also against hate speech (Schieb & Preuss, 2016). If individuals need to see the consequences of behavior to learn its appropriateness, then we have that:

Hypothesis 2 *Observing verbal sanctions to previous examples of hate speech strongly signals the existence of the injunctive norm and will lead to less hostile content.*

1.2 The Experiment

1.2.1 Experimental Design

To test the hypotheses, we designed an online forum where participants could discuss current social topics. The online forum is designed to resemble an Internet forum.[10] Participants were invited to join the conversations and leave comments on topics portrayed in pictures. We collected their comments and later analyzed them.

Pictures and topics were selected in a pre-experimental online survey (N=90) from a list of 10 different social topics and 200 pictures.[11] We chose topics and pictures identified as controversial in the survey to ensure that all topics were, to some extent, subject to public debate. In total, nine

[10] The forum was created using Otree (Chen et al., 2016), a software for economics experiments.

[11] Pictures were previously collected from online media. We used Twitter and Google images to collect the pictures, using a set of keywords (for the list of keywords, see section C.1).

pictures illustrating four topics were selected: three pictures on feminism, two pictures on LGBT rights, three pictures on refugees and multiculturalism, and one picture representing poverty. In a pre-experimental session, we made our forum available online and collected comments on the pictures. A team of three external raters classified a pool of 840 comments based on their level of hate speech into three categories: neutral, friendly, and hostile. The comments were later used to create the experimental conditions.

To test the effectiveness of censoring and counter-speaking, we modified the comments thread in the discussion forum among treatments, while maintaining the order in which the pictures were presented. All comments used to create the experiment came from the pre-experimental session, including the comments used as peer-sanctions in the counter-speaking treatment. The complete experiment time line is described in section C.1.[12]

1.2.2 Experimental Treatments

We implemented four different treatments: baseline, censored, extremely censored, and counter-speaking. The treatments vary in the comments composing the discussion threads in the forum. In the baseline condition, participants could see a balanced mix of two friendly, two neutral, and two hostile comments.

Censored Conditions We implemented two versions of the censored conditions to test Hypotheses 1a and 1b, respectively: censored and extremely censored. Both conditions were designed to highlight a descriptive norm against hate expression. In the censored condition, we deleted prior hate content and presented participants only with friendly and neutral comments. In the extremely censored condition, we presented only friendly comments. Information on whether comments had been deleted was not displayed.

Counter-speaking Condition In the counter-speaking condition, the hostile comments were presented with replies highlighting the unacceptability of hostile opinions (e.g., "[user] this is a prejudiced judgment"). The replies are verbal sanctions that make an injunctive norm salient. The replies were collected from participants in a pre-experimental session. A total of 117 verbal sanctions to hostile comments were collected.

Treatment	**Summary of forum content**
Baseline	6 comments: 2 friendly, 2 neutral, and 2 hostile
Censored	4 comments: 2 friendly, and 2 neutral
Extremely censored	3 comments: all friendly
Counter-speaking	6 comments: 1 friendly, 1 neutral, and 2 sanctions

Table 1.1: Summary of the content of the online forum in the different treatments.

[12] The survey was conducted in April 2016. The pre-experimental collection of the comments was carried out on June 2016. Finally, the experiment was conducted on 4 and 5 August 2016.

The exact comments and the order of appearance in the discussion were automatically selected from a database for each participant in the experiment.[13] Table 3.1 summarizes the number of comments in the different experimental conditions.

1.2.3 Data Collection

The experiment was conducted entirely online and participants were recruited from a crowdsourcing Internet marketplace.[14] The experiment was conducted entirely in German and the sample was restricted to residents in Germany. Although we did not directly ask participants for their demographic characteristics, the subjects were selected from a population with the characteristics depicted in Table A2. The sample is obviously more diverse than the traditional convenience sample of students.

Gender	
Women	55%
Age	
18-24	28%
25-34	42%
35-44	17%
>45	13%
Employment status	
Student	29%
Employee	26%
Self-employed	15%
Other	20%
N.S	10%

Table 1.2: Sociodemographic characteristics of the population from where subjects were recruited.

Participants were compensated with a fixed amount of three euros. To avoid demand effects, the participants were told that they were taking part in an experimental study, but not told the

[13] As shown in Table 3.1, the number of comments differs by treatment. One might argue that the different number of comments could have an impact on the level of hostility and, therefore, be a confounder of the treatment effect. For example, fewer comments might discourage participants to comment. The decision to vary the number of comments was a design choice to keep the amount of friendly content more or less equal between treatments, and to avoid suspicious designs. Nevertheless, no traces of discouragement effect due to a low number of comments were found, since the number of invalid comments was evenly distributed among treatments. An information reduction effect does not seem probable, since the condition with the less displayed information is not the one with less participant-generated hostile content.

[14] We recruited participants from Clickworker (`www.clickworker.com`). The advantage of using this platform was that we could prevent subjects from participating more than once in our experiment, a widespread problem of online experiments. The disadvantage was that using online workers in the experiment could also raise concerns about the external validity of the experiment. Findings in this experiment may not generalize to everyone in every context, but we argue that they generalize to users in online forums because, despite being paid, participants remained anonymous and could abandon the experiment at any moment, which makes their comments voluntary. Since the comments where voluntary, we believe that their motivation to express hate should not differ from motivations of the general population of online forum commenters and, if they do, these differences are constant among treatments. Furthermore, using online workers is, if anything, underestimates the prevalence of hate speech. Since we are not interested in prevalence estimates, but in the effectiveness of our treatments, we do not consider this a problem.

purpose of the experiment. Links were posted in the recruiting platform, and upon acceptance participants were redirected to our own online forum. Participants were randomly allocated between the conditions and asked to join the discussion forums. Each participant was then showed the introduction page explaining the nature of the task. Participants could abandon the experiment at every stage of the experiment just by closing the browser. At the beginning of the experiment, they were given a randomly generated neutral user name and avatar. Every participant was consecutively presented with the nine discussions and asked to leave a comment at the bottom of each thread. Giving a comment was mandatory in each of the nine discussions. Navigation throughout the online forum was always forward. It was not possible to go back to previous discussions once a comment had been sent. When the experiment was completed the participants were given a code to claim the payment for participating in the experiment.

Treatment	**Subjects**	**Comments**
Baseline	47	375
Counter-speaking	45	373
Censored	46	377
Extremely censored	42	344
Total	180	1469

Table 1.3: Number of participants and valid comments in each treatment.

A total of 180 participants were recruited to take part in the forum. Participants spent an average of ten minutes in the forum. We collected a total of 1585 comments, of which 116 were invalid.[15] The comments were evenly distributed among the pictures, with a maximum of 180 and a minimum of 174 per picture. Participants could not see what other participants immediately before them had commented, but only the comments we had previously selected to create the different conditions. This ensured that individual observations were independent.

1.2.4 Measurement of Variables and Operationalization

We evaluated the comments in two ways: we assigned them a score, using a 9-point scale measuring hostility; and we identified those that were clear violations of an anti-hate speech norm. The score tries to encompass a broad definition of hate speech in terms of “tolerance, civility, and respect to others” (Gagliardone et al., 2015, p. 15). This measure is related to the notion of hate speech based on social norms of polite expression. The second measure refers more to a notion of hate speech similar to those found in legislation and international agreements, such as the policy recommendations of the European Commission against Racism and Intolerance (2016), which name the most egregious forms of hate speech.

[15] The number of invalid comments is evenly distributed among the treatments: 33 in the baseline, 29 in the censored, 22 in the extremely censored, and 32 in the counter-speaking treatment. A comment is considered invalid when it is unintelligible. Participants were asked to comment on 9 different threads. We ran the analysis excluding the comments of those who failed to leave 9 comments (N=6), and the results did not change.

Hate Speech Score

The first outcome of interest is the change in the level of hostility displayed by the subjects in the different treatments compared to the baseline group. Thus, the collected comments were classified following a hate speech score by three external raters blind to the experimental conditions. Raters were provided with each comment and the question: "Is the comment friendly or hostile towards the group represented in the picture? (Give a number from 1 to 9 where 1 means very friendly and 9 means very hostile)". Comments with a score of 1 are very friendly in language and express a positive opinion (e.g., Comment 110: "Very brave, I find it great and refreshing. I find despising homosexuals generally bad", user 84. Retrieved from a thread on LGBT rights), whereas a score of 9 normally implies aggression (e.g., Comment 1029: "Gay guys are the last thing I would tolerate, especially in public", User 112. Retrieved from a thread in LGBT rights).

We opted for a continuous measure instead of a binary classification because binary classifications of hate speech have been found not to be very reliable (Ross et al., 2016). Inter-rater reliability of our scale is relatively high (Krippendorf's $\alpha = 0.69$).[16] Thus, we averaged the three scores to construct a hate speech score. The continuous score allows us to study subtle variations and serves as the main variable of interest in the study.

Hate Speech Indicators

We identified various items in the literature that are consistently considered instances of hate speech: 1) contains negative stereotypes, 2) uses racial slurs, 3) Contains words that are insulting, belittling, or diminishing, 4) calls for violence, threat, or discrimination, 5) uses sexual slurs, and 6) sexual orientation/gender used to ridicule or stigmatize. These items are based on guidelines on how to detect online hate speech, published by UNESCO (Gagliardone et al., 2015), as well as the ECRI general policy recommendation on combating hate speech (2016).[17] Comments were labeled as violations of the anti-hate-speech norm if they contain one of the listed indicators. This measure was created with the intention of having a more systematic classification of the norm violations, which could be used for robustness checks.

The two variables measure different things, which can lead to mismatches. Nevertheless, they are closely related and, as the value of the score increases, the probability of a comment being labeled

[16] We computed different measures of inter-rater agreement and reliability such as intra-class correlation (ICC=0.704). We chose Krippendorf's α to assess inter-rater reliability. This measurement is commonly used by researchers in content analysis (Krippendorf, 2004), and it is well suited to handling missing data, as well as specially recommended for cases with more than two raters. The level of agreement differs between the topics. The maximum level of agreement is found in refugees/multiculturalism ($\alpha = 0.71$) and the lowest in LGBT ($\alpha = 0.58$).

[17] "Considering that hate speech is to be understood for the purpose of the present General Policy Recommendation as the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of "race", colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status." (ECRI, 2016, p. 3).

as hateful also increases.[18] The following comment is a typical example of hate content in the forum with a score of 8.66. Comment 159: “Refugee crisis. They can continue walking away from Europe. They are not just war refugees, 90 per cent are nothing but social parasites who can do whatever they want here.” (User 171. Retrieved from a thread on refugees/multiculturalism. The original comment is in German.). The comment was also marked as containing items 1 and 3 by the three raters and therefore classified as a norm violation. More examples of comments can be found in section C.1.

1.3 Data and Results

1.3.1 Average Levels of Hate Speech

We begin this section by analyzing the hate speech score. The mean differences in hate speech score by treatment across topics are displayed in Figure 1.1. The mean hate speech score is reduced in all treatments compared to the baseline treatment (blue line) for all topics except poverty.

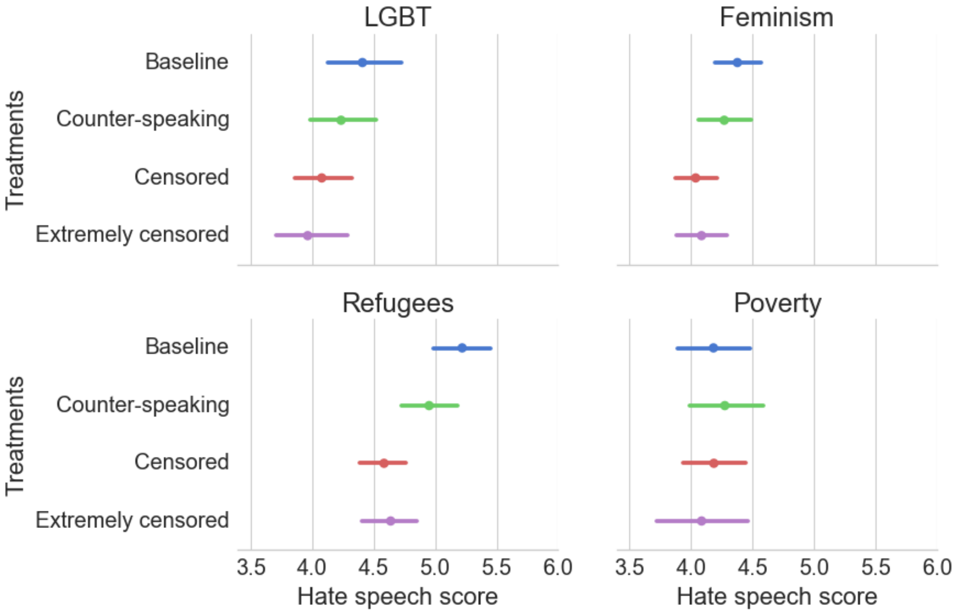


Figure 1.1: Treatment differences in mean hate speech score across topics (Obs=1469). Error bars at 95% confidence interval.

We estimated a random intercept multilevel regression model (Judd, Westfall, & Kenny, 2017) with two random factors, subjects and pictures, and hate speech score as the dependent variable

[18] The predicted probabilities of a comment being classified as hateful by the three raters increases from 0.054 at a score of 6 to 0,98 at a score of 9. Only the 5.5% of comments containing an item from the list has an average score of 5 or lower.

to assess the ability of the treatments to reduce average levels of hostility (see Table 1.4).[19] The main explanatory variables are the treatments (model 1), but we also included terms for the different topics (model 2), and a term for each combination of treatment and topic (model 3). Although the effects of the treatments by topics were not part of our original research question, including them allows us to ensure that the effect is not driven by just a single topic.

We computed the following models:

$$Y_{ijk} = \beta_0 + \beta_1 Treatment_{ij} + u_i + v_j + \epsilon_{ijk} \tag{1.1}$$

$$Y_{ijk} = \beta_0 + \beta_1 Treatment_{ij} + \beta_2 Topic_{ijk} + u_i + v_j + \epsilon_{ijk} \tag{1.2}$$

$$Y_{ijk} = \beta_0 + \beta_1 Topic_{ijk} + \beta_2 (Treatment_{ij} \times Topic_{ijk}) + u_i + v_j + \epsilon_{ijk} \tag{1.3}$$

$$where\ u_i \sim N(0, \sigma_u)\ and\ v_j \sim N(0, \sigma_j) \tag{1.4}$$

The first model in Table 1.4 shows all treatments compared to the baseline condition. The censored and extremely censored conditions significantly reduced hostility by 0.39 and 0.40 scale points, respectively. These results show support for the descriptive norm mechanism suggested in hypotheses 1a and 1b, although extremely censoring does not have an additional effect on the mean score and the magnitude of the reduction is the same for both treatments. In the counter-speaking condition, the score is reduced by 0.14 points compared to the baseline treatment, but this reduction is not significant. These results show no support for hypothesis 2. The second model in Table 1.4 adds topics as predictors, using poverty as the reference category. After controlling for the topics, the effect of the experimental predictors persists. Comments on refugees/multiculturalism threads are more hostile on average (0.61 scale points), while the rest of topics obtained similar levels of hostility to poverty. The following comments, retrieved from a thread on transgender issues, illustrate what a change from 5 to 6 in the score looks like. Comment 268: “Much more important than the question of man, woman or transgender seems to me the question of why anyone goes to the military.” (User 81. Score of 5). Comment 209: “I am confused.” (User 180. Score of 5,33). Comment 317: “I would claim to have chosen the wrong clothes in the wardrobe in the morning.” (User 25. Score of 6). Model 3 shows the effect of the treatments for each topic compared to their respective baseline levels. The treatments’ main effects are deliberately not included in the model. This way the effect of the treatments is shown for each topic specifically. Because poverty is used as the reference category, the intercept represents the estimate for poverty in the baseline treatment. Although the magnitude and significance of the effect differ between topics, the treatments consistently reduce the score as shown by the negative coefficients. In the case of the counter-speaking

[19] Subjects were asked to leave 9 comments in 9 different pictures; hence the comments are clustered both within subjects and pictures. The models with one random level and two random levels were tested using ANOVA. Both random levels are significant. The magnitude of the intra-class correlation (ICC) estimate, i.e., variance accounted for by between-subjects differences, suggests that variability between subjects is very high and should be taken into account in all analysis

treatment, and in line with model 1, this reduction is not significant for any topic. Similarly, none of the terms for the interaction of the treatments with poverty is significant. The treatment effect is larger in threads discussing pictures portraying refugees/multiculturalism, and both censoring and extremely censoring reduce the score in more than half point in this topic. These results should be interpreted with caution, since the effects of the topics are not part of the original research questions and we do not have enough statistical power to test the assumption of a larger effect in threads on refugees/multiculturalism.

Table 1.4: Results from multilevel random models of hate speech score

	Model 1	Model 2	Model 3
Main effects			
Constant	4.63 (0.17)**	4.41 (0.35)**	4.20 (0.37)**
Counter-speaking	−0.14 (0.16)	−0.14 (0.16)	
Censored	−0.39 (0.15)*	−0.39 (0.15)*	
Extremely censored	−0.40 (0.16)*	−0.40 (0.16)*	
LGTB		−0.00 (0.41)	0.22 (0.44)
Refugees/Multiculturality		0.61 (0.39)	0.97 (0.41)*
Feminism		0.03 (0.39)	0.19 (0.41)
Interaction effects			
Poverty×Counter-speaking			0.07 (0.24)
LGTB×Counter-speaking			−0.16 (0.20)
Refugees×Counter-speaking			−0.26 (0.19)
Feminism×Counter-speaking			−0.10 (0.18)
Poverty×Censored			−0.02 (0.24)
LGTB×Censored			−0.33 (0.20)$^{\cdot}$
Refugees×Censored			−0.64 (0.19)**
Feminism×Censored			−0.33 (0.18)$^{\cdot}$
Poverty×Extremely			−0.12 (0.25)
LGBT×Extremely			−0.45 (0.21)*
Refugees×Extremely			−0.60 (0.19)**
Feminism×Extremely			−0.28 (0.19)
Groups:Subjects	180	180	180
Var:Subjects	9	9	9
Groups:Subjects	0.44	0.44	0.44
Var:Subjects	0.15	0.11	0.11
Residual Variance	0.90	0.90	0.90
ICC: Subjects	0.30	0.30	0.30
ICC: Pictures	0.10	0.07	0.07
AIC	4345.59	4347.50	4371.48
BIC	4382.64	4400.42	4472.04
Log Likelihood	-2165.80	-2163.75	-2166.74
Obs	1469	1469	1469

Notes: Linear mixed model fit by REML. Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for models of hate speech score. Model 1 shows main effects of treatments, Model 2 shows main effects of topics, and Model 3 shows the interaction between treatments and topics after controlling for topic main effects. The table lists mean regression coefficient estimates with standard errors in parentheses and p-values calculated based on Satterthwate's approximations. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test

1.3.2 Distribution of the Hate Speech Score

Figure 1.2 displays the distribution of the hate speech score of each treatment compared to the baseline. Extreme comments, both extremely hateful and extremely friendly, are rare, with a majority of comments classified as neutral. The distribution of the hate speech score in the baseline and the counter-speaking conditions are similar. In contrast, in both censored treatments the distributions are skewed to the left, which means that comments were less hostile on average (for both treatments compared to baseline a Kolmogorov-Smirnov test of equality of the distributions yields $P < 0.001$).

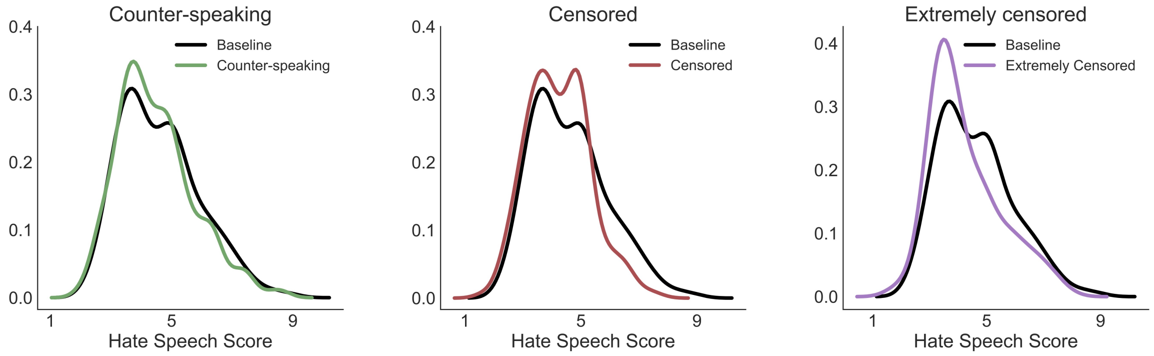


Figure 1.2: Density estimates of average hate speech score in the counter-speaking treatment (green), the censored condition (red), and the extremely censored condition (purple), compared to baseline treatment (black). Size of bins selected using Freedman–Diaconis rule. All treatments are compared to the baseline group (N=1469). The graph depicts the 1 to 9 score scale: scores on the left correspond to friendly speech, while scores on the right correspond to more hostile language.

Next, we analyze displays of hostile comments. We define hostile comments as those with a 7 or more in the score. Hostile comments are relatively uncommon (N=61). Of the comments in baseline treatment, 24 were hostile (5.88%), compared to 4 comments in the censored condition (0.99%). The reduction is significant [$\chi^2(1, 814) = 14.94, P < 0.001$].[20] There are 18 hostile comments in the counter-speaking treatment (4.44%) and 15 in the extremely censored treatment (4.10%). None of them significantly reduces extremely hostile comments. Similar results are obtained using a quantile regression. We computed the treatment effect in the .75, .90, .95, and .99 quantiles of the hate score distribution. The censored condition significantly reduces hate in the .75, .90, .95, and .99 quantiles, whereas a significant effect of the extremely censored condition is found only at the 0.75 and 0.99 quantiles. The treatment effect of the censored condition is also larger in the higher quantiles than the mean effect, e.g., a reduction of 1.33 scores points in the 0.99 quantile compared to the baseline (see section C.2). Of the total number of participants, 33 left a comment that was classified as hostile. Most participants that left a hostile comment left 1 or 2 in total. The maximum number of hostile comments left by one unique participant is

[20] Our findings are robust to the different inference methods as displayed in Table A2 in section C.2. The effect of the censored treatment is robust to the removal of influential individuals. These analyses are available upon request

7. Table 1.5 shows the distribution of the total number of hostile comments made by participants that made at least one hostile comment.

In the extremely censored condition, there are more comments with hostile scores than in the censored condition [$\chi^2(1, 721) = 7.63, P < 0.01$]; even though participants shift their tone (from neutral to slightly friendly), there is an upturn of hostile language compared to the censored condition. This upturn effect is not robust if we take into account the nested structure of the comments, that is, it disappears when we analyze the distribution of hostile comments from the participants' perspective (see section C.2).

Number of hostile comments per participant

Treatments	n=0	n=1	n=2	n=3	n=4	n=5	n=6	n=7
Base	36	6	0	4	0	0	1	0
Counter-speaking	34	7	3	0	0	1	0	0
Censored	43	2	1	0	0	0	0	0
Extremely Censored	34	6	1	0	0	0	0	1

Table 1.5: Distribution of total number of hostile comments per participant

1.3.3 Analysis of the Norm Violations

In addition to the hate score, we analyzed comments that were classified as a norm violation according to our hate speech indicators, that is, comments that are regarded as uncivil.[21] In the analysis only, comments that were classified as a norm violation by two or three of the raters (majority rule) were used (N=147).

We tested for differences in the frequency of hate comments among the different treatments, using a multilevel logistic model with a random effect for participants (Table 1.6).[22]

The predicted probability of observing a norm violation in the baseline treatment is 0.16. This probability is reduced to 0.06 in the censored condition and 0.10 in the extremely censored condition. Nevertheless, only the censored condition presents a significant reduction of the number of norm violations compared to the baseline condition. Again, we find support for the descriptive mechanism suggested in Hypothesis 1a. We found no significant differences in the number of norm violations by topic. Table 1.7 shows the distribution of uncivil comments made by participants that made at least one.

[21] The agreement between the raters is low (Krippendorf's $\alpha = 0.40$), (Ross et al., 2016) found that inter-rater agreement for binary classification of tweets as hateful or not hateful is very low

[22] Our results are also robust to different testing methods (See Table A3 in section C.2), and to the removal of the most influential individuals (analyses available upon request).

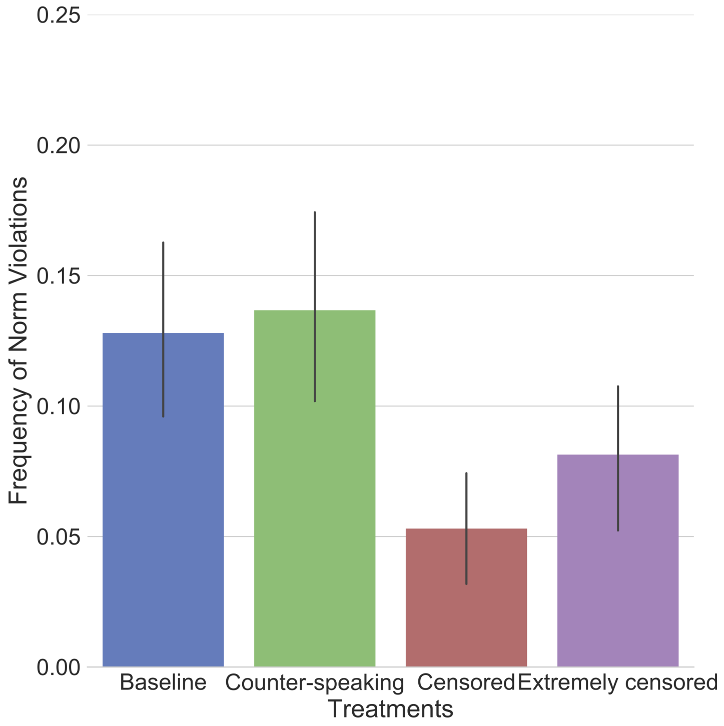


Figure 1.3: Proportion of comments that were classified as hate speech across treatments (N=1469). Error bars at 95% CI.

Table 1.6: Results from a multilevel logistic model of the probability of a norm-violation

	Model (1)
Main effects	
Constant	−2.53 (0.30)***
Counter-speaking	0.19 (0.38)
Censored	−1.00 (0.42)*
Extremely censored	−0.50 (0.41)
Random Pars	
Groups: Subjects	180
Var: Subjects	1.57
Groups: Pictures	9
Var: Pictures	0.05
AIC	892.46
BIC	924.22
Obs.	1469

Notes: Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for models of norm violations. The table lists logistic regression coefficient estimates with standard errors in parentheses and p-values calculated based on Satterthwaite's approximations. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test

	Total number of Uncivil Comments per Participant							
Treatments	**n=0**	**n=1**	**n=2**	**n=3**	**n=4**	**n=5**	**n=6**	**n=7**
Baseline	26	10	3	3	3	1	1	0
Counter-speaking	20	10	9	4	1	0	0	1
Censored	33	7	5	1	0	0	0	0
Extremely Censored	23	16	0	1	1	1	0	0

Table 1.7: Distribution of number of norm-violations per participant

1.3.4 Limitations of the Study

There are three potential limitations concerning the external validity, the generalizability of our results, and the statistical inference. First, the static nature of the forum prevents people from engaging in repeated interaction, which departs from normal dynamics in online forums. The lack of interaction can be important to explain the failure of the injunctive mechanism to reduce hostility in the forum. If, for instance, the commenter expects their comments to be counter-commented, they might be more hesitant to post hateful content.

Second, our sample of participants is limited to online workers, whose characteristics may vary from the general population, thus limiting generalizability of the results. Online workers might differ from the average user of the Internet in their inclination to post hateful comments. This point is not limited to online labors markets, but also applies, for example, to left-leaning or right-leaning websites. Because participation in the experiment was anonymous and voluntary, we believe that their motivation to express hostility should not differ from motivations of the general population of online forum commenters. Although we have no reason to assume that the particular treatment effects are qualitatively changed by our sampling strategy, the results in this paper should not be interpreted as prevalence estimates of hate speech. We acknowledge that our treatments might have different effects for different people, i.e., no effect for those with a strong ideological leaning. Our data does not allow us to test this hypothesis. From a practical point of view, this would mean that providers of online platforms would have to apply very different policies for different people, which is normally not the case.

Finally, the limited size of the sample poses some problems to the statistical inference, especially when analyzing rare events as hostile comments. A bigger sample would be helpful, and this could be collected, for instance, from existing websites and social platforms. This data would be observational, with the endogeneity problem that goes along with it. By contrast, our data is collected in a controlled experimental environment, and therefore allows for a proper identification of the treatment effect.

1.4 Conclusion

The widespread use of social media has become a reality in an ever more densely connected world. One of the biggest social challenges regarding social media is to tackle the hateful speech present in online discussions because it can prevent minority groups from joining conversations and expressing their opinions. We introduced an original randomized experiment to test whether reducing the perceived acceptability of hostility decreases the prevalence of online hate speech. We used specific interventions that used previous actions of others as a source of norm-relevant information.

In line with our first hypothesis (H 1a), we find that moderately censoring hate content reduces the occurrence of further hate comments. Participants were less likely to make use of hostile speech when they were presented with an environment in which previous extreme hate content had been censored. Our results suggest that people respond to cues, in the online context, which signal social acceptability. They do so even when others are unknown and participants remain anonymous, and in the absence of direct punishment. The empirical results do not fully support our second hypothesis (H1b). The general tone of the comments became friendlier by applying an extreme version of censoring, but the effect is similar to moderately censoring. Moreover, there are significantly more extreme hostile comments in the extreme censoring conditions then in the censored condition, which hints at a polarization of opinions. Our intuition is that this result might indicate either a reactance effect [23] or an increase in normative ambiguity.

Both experimental manipulations were more effective for the threads on refugees or multiculturalism. This differential of effects is related to the level of public debate on the topic: a lot of public debate of a topic may increase the salience of the norm (e.g., people might have previously observed that extreme opinions have been sanctioned). However, the censored conditions also allows for a potential competing mechanisms such as mimicking previous comments. Nevertheless, the high prevalence of hate comments in the extremely censored condition indicates that people do not merely imitate observed behavior, but they interpret the actions of others as contextual cues.

The counter-speaking condition meant to test Hypothesis 2 had no significant effect. A potential explanation is that the verbal sanctions, i.e., the counter-comments, might add ambiguity about the norm (the verbal sanctions are, essentially, hostile comments themselves) by putting descriptive and injunctive information in conflict. In ambivalent situations like this, in which more than one norm may apply, individuals may interpret the situation in a way that favors them (Xiao & Bicchieri, 2010; Bicchieri & Chavez, 2013; Winter, Rauhut, & Miller, 2017).

Our findings contribute to the sociological literature in social norms by raising the question of whether descriptive norms might, in some settings, be more effective than sanctions at preventing antisocial behavior. Our results suggest that normative behavior in online conversations might, in fact, be motivated by descriptive norms rather than injunctive norms. This is a surprising effect, given the results from previous research on social norms that pointed to a large effect of sanctions on normative behavior (Heckathorn, 1988; Coleman, 1990c; Voss, 2001). Lab experiments on social norms show similar findings. For example, Fehr and Gächter (2000) conclude

[23] Reactance appears when an individual facing a persuasive message reacts by engaging in the proscribed behavior (Burgoon, Alvaro, Grandpre, & Voulodakis, 2002).

that punishment is far more effective than mere suggestions of maintaining a cooperation norm in a lab experiment. Furthermore, when the effects of injunctive and descriptive norms have been tested together, they do not significantly differ from each other (Krupka & Weber, 2013). Nevertheless, this result should be taken with precaution, since of injunctive information might be weakened by the lack of interaction.

The experimental nature of the study allows us to exclude potential confounding factors that can substantially bias the analysis of observational data. The randomization of subjects between conditions eliminates selection effects, e.g., hateful commenters joining only hateful discussions, and the anonymity in the forum prevent the occurrence of group identification processes. This study overcomes the identification problems that often arise from estimating the effect of normative influence.

This project is a step forward in the empirical research on online hate speech. First, we show that the observed pattern of behavior act as a situational normative cue in online environments. Second, our findings point to a larger effect of descriptive norms — defined as frequent behavior — on reducing hate speech. Finally, we provide a reliable empirical test of censoring and counter-speaking as interventions, and show that moderately censoring hate content is sufficient to reduce uncivil comments. We believe the results in this study can support the design of online platforms that help reduce the incidence of hate speech in cases where it is undesirable and maintain an open online environment.

Nevertheless, we do not have data on the obvious tradeoff between censoring and free speech; hence, our paper does not represent a position on whether censoring hate content is necessarily socially beneficial. We consider that a social norm intervention (e.g., Tankard & Paluck, 2016) is a good approach to address online hate speech, whose presence is not necessarily considered unlawful, but often regarded as undesired.

Chapter 2

Uncovering Hidden Opinions: The Effect of Social Acceptability on Disclosure of Anti-immigrant Views

This chapter builds on a central idea from the previous experiment: whether certain opinions are perceived as socially accepted depends on the predominant pattern of behavior in the social context, that is, the descriptive social norm (Bicchieri, 2016, 2006). A social norm is defined as the preference to conform conditionally to the behavior of the group if we think enough people will conform, too. The conditional conformity assumption is critical to understand how norms can vary from context to context depending on social cues. When enough people commit to the norm against prejudice, people tend to refrain from expressing those views (Álvarez-Benjumea & Winter, 2018; Cheng et al., 2017; Munger, 2016). On the contrary, in environments where the use of hate speech is normalized, people are more willing to disclose previously unacceptable views, such as xenophobic attitudes (e.g., Blanchard et al., 1994; Crandall et al., 2002; Crandall, Miller, & White, 2018; Álvarez-Benjumea & Winter, 2018).

Observing an increasing number of norm transgressions should increase the likelihood of subsequent norm violations because the behavior will be perceived as acceptable in that context. In the case of the expression of prejudice, observing instances of hate speech increases the likelihood of individuals doing so (Álvarez-Benjumea & Winter, 2018; Blanchard et al., 1994). However, it remains unclear how the dynamics actually work. Different mechanisms compete with each other. We take the idea that people use the behavior of others as a reference for the socially accepted behavior as a starting point, and set out to study the dynamics of how the change happens. We assume that an increasing number of norm transgressions will increase the likelihood of subsequent norm violations and elaborate two different mechanisms of how the normative change from anti-hate to prejudice might happen, that is, how the norm breaks down. The first mechanism poses that the change will be linear, that is, norm transgressions steadily increase with observed norm violations; the second, that behavioral reactions will happen only after a certain threshold.

To analyze the breakdown of the norm, the experiment focuses on behavioral reactions. As a behavioral measure, we use donations to a pro- and an anti-immigrant organization: *Alternative for Germany* (AfD) and Pro Asyl. We use these organizations because of their straightforward connection to immigration issues. Particularly in the case of the AfD. After decades of growing ethnic diversity, the integration of immigrants and multiculturalism have become salient political issues. In Europe, narratives blaming ethnic diversity and growing immigration rates for the different political anxieties of the population have become normal. New right-wing populist parties were built or strengthened upon these new narratives. Many of these parties saw a spectacular increase in support, which put them in the government of several European countries. The anti-immigration party AfD, founded in 2013, rose to almost 13% in the 2017 elections. In past years, there has been an increase in votes for parties of the Radical Right Parties (henceforth RRPs).[1]

These parties openly defy norms of political correctness such as the anti-hate norm. The present-day populists, in fact, present themselves as fighters against political correctness (Mudde, 2004; Hirvonen, 2013). Today, a majority of the public opinion still understands the support for RRPs as breaking the anti-hate norm because of its close links to hate speech, anti-immigrant views, etc. As a result, parts of the population, such as women, avoid expressing public support for these parties because of their reputation (Harteveld & Ivarsflaten, 2018). As a very well-known RRP in Germany, the AfD serves as a counternormative option because of its relation with hateful discourses and openly anti-immigrant agenda, which has been described as "bordering outright xenophobia" (Cantoni, Hagemeister, & Westcott, 2017, p. 6).

This paper examines the dynamic of normative change, particularly i) how observing an increasing number of norm transgressions prompts individuals to violate the anti-hate norm, and ii) how reactions to a breakdown of the norm differ conditional on gender. Based on the literature on social norms of prejudice expression, we hypothesize that the perception of increasing social acceptability of prejudice should in general increase its public expression. We introduce two different hypotheses of how the dynamic could look like based on the current theory, a linear and a threshold hypothesis, and test them both. We also test whether the extent of normative influence depends on the gender of the individual. Gender has been found to play an important role in the regulation of social desirability and motivation to control prejudice (Ratcliff et al., 2006), rendering women less likely to support not socially accepted parties and discourses, such as RRPs (e.g., Harteveld & Ivarsflaten, 2016, 2018; Blinder et al., 2013; Immerzeel, Coffé, & Van der Lippe, 2015). Normative change is measured using donations to either an anti-immigrant or a pro-immigrant organization: AfD and Pro Asyl.

We designed an online field experiment (N=2283), in which participants were invited to take

[1] These parties include the Freedom Party of Austria (FPÖ), Marine Le Pen's National Front (FN) in France, Viktor Orbán in Hungary, the UKIP in the UK, and Donald Trump in the U.S., among others (Norris & Inglehart, 2019; Rydgren, 2018). See also"Europe and nationalism: A country-by-country guide", retrieved from `www.bbc.com/news/world-europe-36130006`, last accessed March, 2019

part in an online forum discussing refugees and immigration issues. The social acceptability of expressing prejudice is manipulated by increasing the proportion of comments that contained xenophobic opinions, i.e., norm transgressions. In the exposed group, the number of these comments steadily increases with each consecutive forum page, whereas in the unexposed group there are no hateful comments (the tone of the comments remains stable across forum pages). Participants are given the possibility to make a donation of 1 euro to a randomly drawn organization that could either be anti- or pro-immigration: AfD or Pro Asyl. The time in which the donation question appears is randomized, so participants were asked to make the decision at different stages of the forum. The treatment conditions thus vary along three dimensions: i) the type of organization, ii) the number of comments the participant sees in the online forum before the decision, and iii) the fraction of those comments that are violations of the anti-hate norm. Across systematic variations of this, we measure how the number of observed norm transgressions influences the decision to also violate donate.

The empirical results show that, overall, people are more willing to donate to the pro-immigration organization than to the anti-immigration and that women are particularly reluctant to donate to the anti-immigrant organization. Results also show that women reduced, even more, their donations to this organization when the anti-immigrant comments introduced normative concerns. We explain this result as women displaying greater social desirability bias and more willingness to follow the social norm against prejudice. Results in this paper can help explain how changes in the normativity of openly expressing xenophobic views can impact the success of right-wing populist parties, and how the anti-hate norm could potentially prevent them from gaining further support.

2.1 The Dynamics of Normative Change

Whether certain opinions are perceived as socially accepted depends on the pattern of behavior predominant in the social context; therefore, people are more willing to support certain ideas once they become normalized or approved in their context. First, normalized behavior is less likely to bear punishment (Bicchieri, 2006, p. 11), and second, people prefer to be well-considered by others (Cialdini & Goldstein, 2004). Social norms are expectations of how many others will follow the norm, that is, expectations about others' behavior (Bicchieri, 2016, 2006). Accordingly, changes in these expectations should induce changes in behavior. As a result, if the number of people who openly express xenophobic opinions grows, people should feel less pressure to inhibit certain opinions or even feel compel to adopt them.

The individual behavior of others serves as a powerful normative cue that signals the status of the norm in that particular context (Cialdini & Goldstein, 2004; Cialdini, Kallgren, & Reno, 1991). Observing enough examples of norm transgressions might lead people to think that the behavior is not proscribed in that specific situation at that specific time, i.e., weakens the normative expectation. As a result, the previously established norm might unravel. In the case of the expression of prejudice, observing instances of hate speech in the context increases the likelihood of individuals doing so. For example, observing instances of hate speech in an online forum increases the likelihood of new participants using hate speech (Álvarez-Benjumea & Winter, 2018; Cheng et al., 2017), and hearing others endorse racist views prompted participants to do so as well (Blanchard et al., 1994; Crandall et al., 2002; Crandall & Stangor, 2005; Stangor et al., 2001).[2] The *conditional compliance* assumption of social norms is crucial for the breakdown of the anti-hate norm because it means that the level of conformity should increase with the number of others who visibly comply — or do not comply — with the norm. A higher number of xenophobic comments might lead people to believe that the behavior is acceptable in that context. It follows that an increasing number of norm transgressions will increase the likelihood of subsequent norm violations. However, it remains unclear how the dynamics actually work. How many norm transgressions are *enough* to signal an unraveling norm?

Based on the conditional preference for conformity to norms, to observe a behavioral change most people have to simultaneously believe that most others will change their behavior, such as adopting or abandoning a norm (Bicchieri, 2016, pp. 116-118). Conformity to norms is a complex behavior, that is, it normally requires people to expect that a majority will approve of the behavior before adopting it (Centola, 2018; Centola & Macy, 2007). Generally, any behavior that could be costly if coordination failed is supposed to follow this pattern. Conformity to social norms, adoption of new technology, joining a social movement, or human cooperation follow a pattern of complex contagion.[3] Violating the anti-hate norm carries an inherent ambiguity about its social acceptability because of the inherent risks of being punished or rejected (Dimant, 2018). Until the norm has been challenged by —enough— peers, uncertainty remains high.

In this case, the increasing number of norm transgressions should produce behavioral effects only when most of the people believe that a majority will approve of the behavior. When that threshold is met, people will change their expectations about the norm and therefore change their behavior accordingly. Furthermore, the first norm transgressions might just generate a focusing effect whereby attention is drawn to the norm. Drawing the attention to the norm can impact the behavior by making people think within the normative framework (Krupka & Weber, 2009) and counter-intuitively elicit norm-consistent behavior (Cialdini et al., 1990). The combination of the conditional conformity and the focusing effect will result in a threshold response to the increasing norm violations, such as the one depicted in the left plot in Figure 2.1, in which behavior takes place after the majority threshold of has been met.

For practical reasons, we define most of the others as more than 50% of others and arrive to the first hypothesis:

Hypothesis 1 *Support for anti-immigration views, as measured by the donations, increases only after at least half of the comments are norm transgressions.*

[2] Indeed, the reverse is also true. Providing people with the expectation that others will follow the anti-prejudice norm make then avoid prejudiced language (Paluck & Green, 2009). It works in both directions.

[3] Complex behavioral contagion is the contagion that requires contact with "multiple adopters" (Centola, 2018, p. 7). In contrast, simple contagion, like a rumor going viral, only requires one contact to spread.

Nevertheless, an important aspect of conformity is that different people have different thresholds to conform (see Sliwka, 2007). Different individuals react differently or react after a certain threshold of transgressions has been met. They might have different previous beliefs about the proportion of followers or different thresholds for how many transgressions they need to observe to do so themselves (Bicchieri & Dimant, 2019; Bicchieri & Funcke, 2018). It might happen that people who hold xenophobic attitudes may reveal them just after hearing someone else doing so. Other people will only join if at least half of the others do, and still others will need to perceive a majority. If the thresholds are randomly distributed in the population, we might assume that the effect of an increasing number of anti-immigrant comments will have a linear effect of behavior. Then, the relation will be similar to the one depicted on the right plot in Figure 2.1. This leads to the second hypothesis, which predicts a linear relationship between the number of norm transgressions, i.e., the anti-immigrant comments and the behavioral change:

Hypothesis 2 *Support for anti-immigration views, as measured by the donations, increases steadily as the number of anti-immigrant comments in the forum increases.*

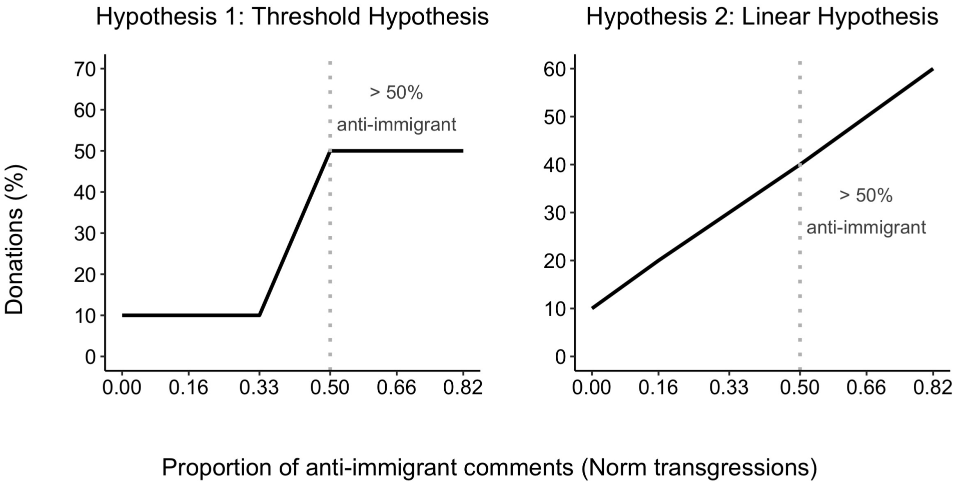


Figure 2.1: Graphical depiction of the relation between the proportion of anti-immigrant comments (norm transgressions) and the expected behavioral response. The plot on the left correspond to hypothesis 1 of a threshold relation. The plot on the right depicts hypothesis 2 of a linear relation.

Each of the hypotheses corresponds to possible ways participants might react to the increasing norm violations. The underlying general idea is that the spreading of the norm violations will depend on the distribution of thresholds in the population. Furthermore, the dynamic might also change because of personal characteristics that produce changes in individual susceptibility to the norm (Walker, Sinclair, & MacArthur, 2015), or prior beliefs about the actual proportion of norm followers. The third hypothesis tests the effect conditional on gender.

2.1.1 Gender Differences in the Susceptibility to the Descriptive Norm

One of the earliest findings of support for populist radical right-wing parties is that they consistently attract more men than women (e.g., Betz, 1994; Coffé, 2018; Givens, 2004), a relationship that persists even after controlling for socioeconomic and political characteristics (Immerzeel et al., 2015). However, the ideological positions of men and women with respect to attitudes at the core of the RRPs politics do not vary as much. Women are as likely as men to hold anti-immigrant views and nativist concerns, but even those women who commune with these values are less likely to vote for the RRPs than men (Harteveld, Van Der Brug, Dahlberg, & Kokkonen, 2015). One explanation of this gap takes into account the existence of the anti-racism norm. This account explains how voting for these political parties also brings about normative concerns because of the openly xenophobic agendas of these political parties. This explanation builds on work from Blinder et al. (2013); Harteveld and Ivarsflaten (2016), who show how the gender gap in votes to RRPs could be explained by women being more motivated to control prejudice.

The motivation to control prejudice is the conscious process of acting according to the anti-hate norm. It has its source in a motivation to commit to external — normative — pressure. This renders personal attitudes and normative behavior independent; Therefore, even those who hold anti-egalitarian and xenophobic attitudes could comply with the norm. Indeed, women have been found to display a larger motivation to control prejudice (Ratcliff et al., 2006). Because the motivation to control prejudice is related to normative concerns about the open expression of xenophobic attitudes, normative cues can trigger this concern by making people focus on the norm. A hate comment in a previously normative environment may affect the extent to which people are drawn to focus on the norm. If an individual has a large normative concern or holds a prior belief of the norm being strong, the hateful comment will only serve to remind that individual of the norm, which in turn may prompt them to show norm-consistent behavior.

Hypothesis 3 *Women will show more reluctance to support anti-immigration views, as measured by the donations, and display behavior consistent with the anti-racist norm longer than men.*

2.2 The Experiment

2.2.1 General Procedure

This paper examines the dynamic of normative change, particularly how observing increasing norm transgressions prompts individuals to violate the anti-hate norm. We test two different hypotheses of how the dynamics of normative change could be: the threshold hypothesis (Hypothesis 1) and the linear relation hypothesis (Hypothesis 2). We also test whether the extent of normative influence depends on the gender of the individual (Hypothesis 3). The local context in the experiment is an online forum in which participants are asked to participate and comment on pictures depicting images of refugees and immigrants. A total of 5 pictures were selected (see subsection B.1.2). To assess normative change, we use a behavioral measure: donations to an organization well-known to be either for or against welcoming refugees. Donations only have a signaling value since participants do not incur any personal monetary cost.

The selected organizations are *AfD* and *Pro Asyl*. The selection was made based on two criteria. First, an online questionnaire was implemented to select organizations that were widely known to the population. We provided respondents with a list of nine different organizations with different stated views about immigrants and ethnic integration and asked which ones participants had heard of before. We also ask the respondents to classify the organizations based on whether they believed they were for or against immigration. Second, the organizations were selected based on their relation to the anti-racism norm.[4] The selected anti-immigrant association, the AfD, is known for openly defying the anti-racism norm with its discourse, and has adopted a rhetoric emphasizing topics such as nationalism, anti-immigration policies, and the so-called threat of Islamization (Cantoni et al., 2017).[5]

The perception of local social acceptability of anti-immigrant views is manipulated by altering the fraction of comments that were openly anti-immigration or xenophobic. Because a norm is a group's expected behavior, the behavior is more informative if learned from the people in the same context (see Bicchieri, 2016, 2006). In this case, the relevant others are the participants in the forum. The comments were presented in an increasing manner. All participants took part in five different consecutive forum pages, were asked whether they would like to donate to the selected organization, and asked to complete a short questionnaire on their demographic characteristics at the end. The donation decision appears only once and can appear after any of the forum pages. The decision time was randomized among participants. No other information regarding the social norm, rules, or expected behavior was provided to avoid any conflicting information.

2.2.2 Treatments

The treatment conditions vary along three dimensions: i) the type of organization; ii) the number of comments the participant sees in the online forum before the decision (time of donation: 1 to 5); and iii) whether a fraction of those comments is anti-immigrant (exposed vs. unexposed groups). All variations occur between subjects. The different experimental conditions are the following:

[4] The complete relation of proposed organizations is: Alternative for Germany, Identitäre Bewegung, Bürgerbewegung PRO-NRW, Mensch Mensch Mensch, Fluechtlinge-Willkommen, Pro-Asyl, Gustav-Stresemann-Stiftung, Desiderius-Erasmus-Stiftung, and Amadeu Antonio Stiftung (see Table A1).

[5] Until 2015, the party kept a discourse focused on Greece and the Euro crisis (Cantoni et al., 2017).

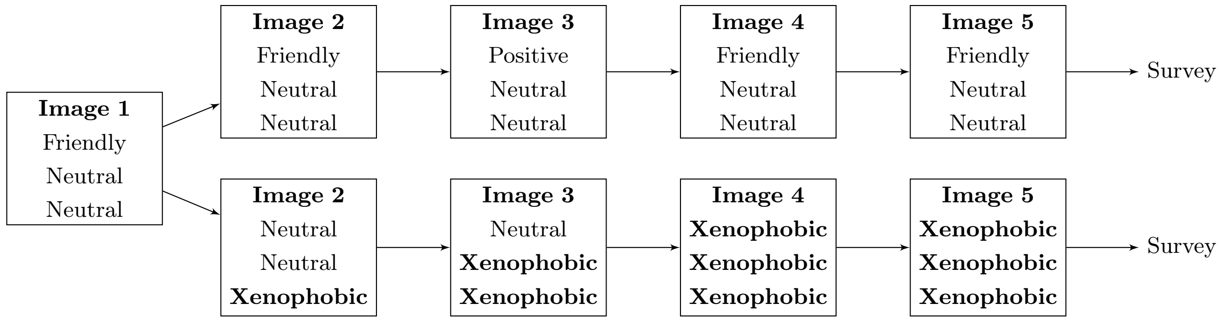


Figure 2.2: Different experimental treatments and their content.

The experiment consists of 2 arms with 5 different times of donation. This is repeated for each association. The arms with no anti-immigrant comments work as the control —unexposed— condition, where no xenophobic messages are shown. Using this setting, the donation rates can be compared, not only to Time 1, but between all times. This setting controls for a possible effect on donations of time spent in the forum. The experiment allows for the comparison of normative versus non-normative environments, as well as the comparison to the baseline.

Forum pages consist of one image on top which depicts refugees, followed by three comments discussing the picture. The number of comments is kept constant. All participants were presented with the 5 consecutive forum pages. The donation decision is exogenous and appears after one of the forum pages picked randomly for each participant. This makes a total of 5 possible treatments (times 1 to 5). Time 1 means that the donation decision was asked after the first forum page, Time 2 after the second, and so forth. The comments vary in their language towards refugees: friendly, neutral and xenophobic. Xenophobic or anti-immigrant comments are violations of an anti-hate norm. The number of anti-immigrant comments provides a measure of the perceived local popularity of anti-immigrant sentiment and strength of the anti-hate norm.

In the upper arm (unexposed group) of the experiment, the number of each type of comment is also kept constant: 1 positive plus 2 neutral comments. The collection and selection of comments are explained in subsubsection 2.2.2. In the lower arm (exposed group), the number of xenophobic comments increases with each forum page. The increase is linear and the fraction of negative comments always increases in 16% (i.e., 1/6) with each new forum page, from 0 comments to a maximum of 66% of comments being hateful.

$$\text{Proportion xenophobic comments:}\quad \begin{array}{ccccccccc} 0 & \longrightarrow & 1/6 & \longrightarrow & 3/9 & \longrightarrow & 6/12 & \longrightarrow & 9/15 \\ & & (16\%) & & (33\%) & & (50\%) & & (66\%) \end{array}$$

Figure 2.3: Proportion of xenophobic comments in each forum page (1-5).

At the end of the experiment, every participant has seen fifteen comments. The association, the treatment, the donation time, and the image order are randomized among participants. The randomization of the order of the images avoids any order effect.[6] After they had participated in the forum, they were asked to complete a brief demographic questionnaire. In the end,

[6] Additionally, the image that the participants see right after the donation decision is also controlled for in the robustness analyses.

participants were told the fraction of others who had decided to donate and a the code for claiming the money earned with the experiment.

Donation Decision

Participants were told that they would be given the possibility to make a donation of 1 euro to a randomly drawn organization that could either be anti- or pro-immigration, to ensure that the participants could not associate the experimenters with a specific political ideology. Some details about the organization and its political principles were provided in the experiment. The instructions are the following:

Treatment A: Pro-social organization:

The FOUNDATION PRO ASYL was founded in 2002. Their goal is to finance refugee and human rights work. The FOUNDATION PRO ASYL implements projects directed to increase refugee integration - such as supporting scholarships for refugees.

Treatment B: Anti-social organization:

AfD (Alternative for Germany) is a right-wing populist, eurocritical political party founded in Germany in 2013. The members of the party call for a structured refugee policy, no further reception of refugees, and a limit for the reception.

In this part of the instructions, the participants were also told that they would be shown the fraction of others who had decided to donate at the end of the task.

The Comments

All comments used to create the forum come from pre-experimental sessions, in which we made the forum available online and collected comments on the pictures. The comments were classified by external online raters using a 1 to 9 scale where 1 is very friendly language, and 9 is very hostile language.[7] Using this score, the comments were categorized as neutral, friendly, and xenophobic. To make sure that participants were attentive to the information displayed in the forum, the number of comments in each forum page was kept low; only three comments were displayed in each forum page. All positive comments were rated with 2 or less in the hate score. Neutral comments were always between 4.5-5.5 in the scale. Finally, xenophobic comments were those with more than 7.5 in the hate score. This sharp classification is meant to reduce ambiguity and the cognitive load of the participants.

[7] The scale answered the following question: Is the comment friendly or hostile towards the group represented in the picture? *(Give a number from 1 to 9 where 1 means very friendly and 9 means very hostile).* The classification of the comments was done by external raters.

2.2.3 Sampling Strategy

The total number of participants is 2283.[8] The unequal assignment of participants to the different conditions is due to the random allocation technique being simple randomization. The allocation was decided upon the beginning of the experiment and the probability of allocation was the same for every treatment. There are no significant differences in the distribution of the number of participants in each cell. Initially, a total of 2399 subjects participated in the experiment; however, 102 participants abandoned the experiment without completing the donation decision, and an additional 14 participants were deleted because their comments were not valid. Unintelligible comments were considered as invalid messages, e.g., gibberish, symbols, and copy-pasted url addresses, or messages that were copy-pasted in all forum pages (70 comments). The deleted comments are shown in section B.1.[9]

Association	**Treatment**	**Donation Time**					**N**
		Time 1	Time 2	Time 3	Time 4	Time 5	
AfD	**Unexposed Group**	122	113	113	119	114	581
	Exposed Group	121	122	113	116	102	574
Pro Asyl	**Unexposed Group**	119	118	123	113	116	589
	Exposed Group	116	107	102	110	104	539
Total Observations		478	460	451	458	436	2283

Table 2.1: Distribution of the total number of participants (N=2283) among the different combinations of association (AfD and Pro Asyl), treatment group (Unexposed and Exposed Groups), and the donation times. The experiment has a total of 20 different possible combinations(2 × 2 × 5).

Overall, 187 (16%) of the participants in the AfD condition and 704 (62%) of the participants in the Pro Asyl condition decided to donate the money. The donation rates are similar for the exposed and unexposed groups on average (see Table 2.2). These figures contrast with the pre-experimental survey, where only 11% and 26% of the participants said they would donate to the AfD or Pro Asyl, respectively, if asked to do so.

Table 2.2: Total number of donations by organizations and experimental condition for all participants.

Association	Condition	Freq (%)
AfD	Unexp.	97 (17%)
AfD	Exp.	89 (15%)
Pro Asyl	Unexp	368 (62%)
Pro Asyl	Exp.	333 (62%)
Total		887 (39%)

[8] Due to the large size of the sampling the experiment run for over a week, starting on the 9th October 2019.

[9] Because this is an online experiment, I have to distinguish between participants who dropped out the experiment (n=102) and those who clicked on the link, but never passed from the welcome page or early abandonment (n=2053).

In total, 891 donations were made: 187 to the AfD and 704 to Pro Asyl. The donations were paid via bank transfer on 18th January 2019. Table 2.3 shows the characteristic of the participants of the experiment based on the post-experimental questionnaire. An extended table describing the allocation of characteristics in the different conditions can be found in section B.1.

Table 2.3: Characteristics of the participants (results from post-experimental questionnaire).

	Freq. (%)
Gender	
Men	1095 (49%)
Women	1091 (49%)
Other	49 (2%)
Employment	
Unemployed	140 (6%)
Employed	1225 (55%)
Not active/retired	102 (5%)
Self-employed	319 (14%)
Student	449 (20%)
Education	
high school	776 (35%)
Bachelor	395 (18%)
Professional Qual.	598 (27%)
Master	375 (17%)
Primary or less	91 (4%)
East/West	
East	329 (15%)
West	1906 (85%)
Inhabitants	
<50.000	859 (38%)
50.000 – 200.000	495 (22%)
200.000 – 500.000	252 (11%)
500.000 – 1.500.000	308 (14%)
>1.500.000	321 (14%)
Age	
18-25	626 (28%)
26-35	795 (36%)
36-45	404 (18%)
46-55	264 (12%)
55+	146 (7%)
Total	2235

2.2.4 Demographic Predictors of Donation

The total number of participants is 2283. The unequal assignment of participants to the different conditions is due to the random allocation technique being simple randomization. The allocation was decided upon at the beginning of the experiment and the probability of allocation was the same for every treatment. A logit regression model with donations as the dependent variable, and the demographic characteristics collected in the post-experimental questionnaire as the predictors were run for each association. The models give a general picture of which characteristics are linked to an increased or decreased likelihood of donation (Figure A2). The demographic questions were chosen based on their plausible effect on donations. The demographic characteristics used to construct the models are gender, employment status, Education, size of the residency, age, and part of Germany (West or East Germany). Particular attention is paid to the role of gender, and whether women on average donate less to the AfD, since a dissimilar effect of gender is crucial for the third hypothesis.

Various demographic characteristics significantly impacted donations to the AfD (the coefficients are depicted in Figure A2). Receiving a pension significantly increased the likelihood of donation when compared to employees ($p = 0.02$). Characteristics that led to a significantly lower probability of donations are: having completed either high school ($p = 0.00$), a Master's degree ($p = 0.03$),[10] and being more than 55 years old ($p = 0.00$). Being a woman reduced the total likelihood of donation to the AfD ($p = 0.00$) by 0.06 points (the predicted probability of donation for men is 0.2, and 0.14 for women). The difference of the predicted probabilities is significant testing the linear combination of the marginal effects ($p = 0.004$).[11]

In the case of donations to Pro Asyl, participants with higher education levels, both Bachelor ($p = 0.01$) and a Master's degree ($p = 0.00$), donate significantly more to Pro Asyl, as well as participants residing in the west part of Germany ($p = 0.00$). Gender, however, does not have a direct effect on the likelihood of donation.

In general, more years of education, and being a woman seem to discourage donations to the AfD. The negative relation between women and donations to the AfD is also found in the pre-experimental survey conducted online with 199 respondents ($p = 0.06$).[12] None of the images has a significant effect on the likelihood of donation, either in the AfD or in the Pro Asyl condition.

[10] Participants who had completed high school had a 0.11 predicted probability of donation compared to the 0.22 predicted probability of participants with professional education, or the 0.20 of participants with only primary school.

[11] The difference in predicted probabilities is computed from a model where every else predictor remains constant. In a model with gender as the only independent variable, the difference is -0.057 ($p = 0.01$)

[12] The results of a logistic model using the pre-experimental survey data are depicted in Figure A3.

2.3 Results

The results are presented in various steps. We first look at the effect of increasing the number of anti-immigrant comments on the dependent variable donation. The likelihood of donation is compared to the baseline likelihood of donation when zero anti-immigrant comments are observed. Furthermore, a regression model with a linear and a quadratic term of the number of comments is also constructed to look for a non-linear increase of donations based on the number of comments observed. In a second step, the probability of donating is compared at each donation time (1 to 5) for each experimental group (Unexposed Group × Exposed Group). Using this setting, the effect of the increasing social acceptability of xenophobic views is tested independently of the effect of time spent within the discussion forum. Hypotheses 1 and 2 are tested using this approach (for ease of analyss, Hypothesis 2 is tested before Hypothesis 1) Finally, we test whether women respond differently to the increasing number of anti-immigrant comments (Hypothesis 3) using similar models where the interaction effect with gender is added. Because of the binary nature of the dependent variable (i.e., donation), results are reported as Average Marginal Effects (AMEs) along with the coefficients from the logit models. For the significance level of the AMEs, both tests of first differences and tests of the second differences for interaction effects are used when necessary (Mize, 2019).[13]

2.3.1 The Effect of Increasing Anti-immigrant Comments

Overall, 62% of the participants donated to Pro Asyl, and 16% to the AfD (Δ=0.46, $p = 0.00$). This difference in percentage of donations between the different organizations persists when only data from the unexposed (Δ=0.46 , $p = 0.00$), and the exposed Group (Δ=0.46, $p = 0.00$) are analyzed. Figure 2.4 shows the frequency of donations for each group (Unexposed and Exposed) at each treatment time (1-5). As described in section 2.2, the number of anti-immigrant comments increases each time in the exposed group, from 0 in Time 1 to 9 comments in Time 5. The proportion of anti-immigrant comments over the total of observed comments increases as follows, from time 1 to time 5: 0, 16%, 33%, 50%, and 66%. No anti-immigrant comments are shown in the unexposed group.

The top plot in Figure 2.4 shows the results for donations to the AfD in the exposed and unexposed group. The results from the logistic model are shown in Table A9, and the average marginal effects in Table A5. There is a slight decrease in donations after the inclusion of 1 anti-immigrant comment in Time 2, although this difference is not significant. After Time 2, a slight tendency to increase in the probability of donating to the AfD is observed in both the unexposed and the exposed group. In the exposed group, the predicted probability of donation is higher than in the baseline only in Time 4 (when 6 anti-immigrant comments, 50% of total

[13] The effect of the anti-immigrant comments should be the opposite for the AfD and Pro Asyl donations. In the case of the AfD, the perception of the increasing social acceptability of anti-immigrant views should increase the number of donations. In the case of Pro Asyl, the increasing anti-immigrant comments should decrease the frequency of donations.

comments, have been observed). The probability of donating decreases again in Time 5 (after 9 anti-immigrant comments, 66% of the total, have been observed). This decrease in Time 5 is not found in the unexposed group, in which the probability of donating increases in comparison with Time 4.

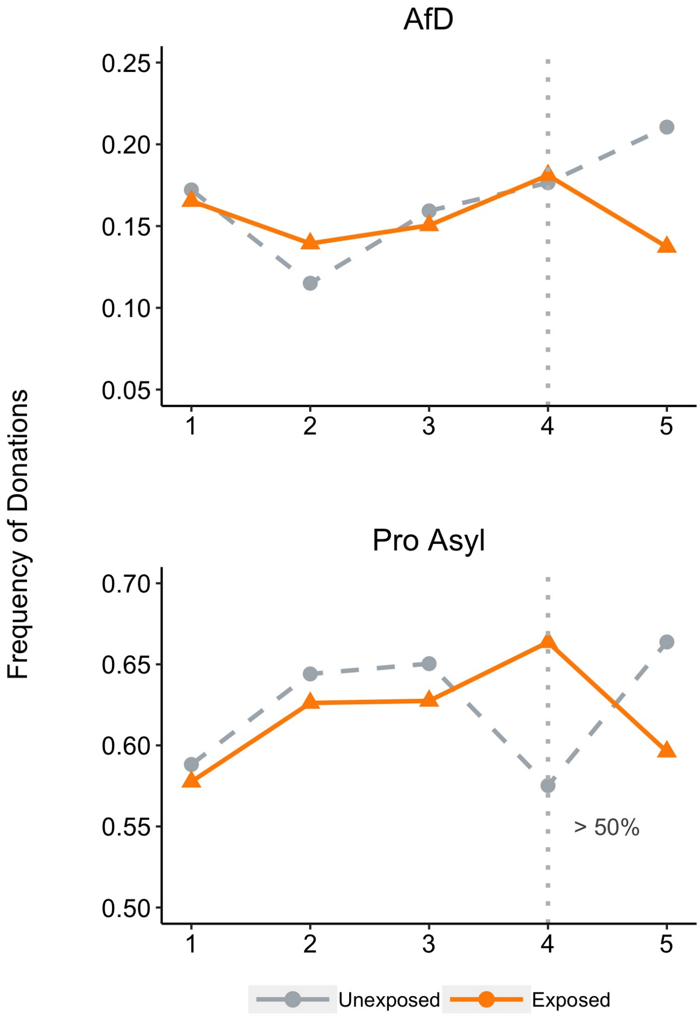


Figure 2.4: Frequency of donation in every stage of the forum for AfD (upper plot) and Pro Asyl (bottom plot). The Orange line represents online forum pages in the exposed group (number of anti-immigrant comments increasing), compared to the unexposed Group (dashed grey line). From time 4, the proportion of anti-immigrant comments is 50% or more, as shown by the dotted line. The number of observations is 2283.

In order to test hypothesis 1, we now look at the regression model with a quadratic term and at the effect of having observed more than half of the comments as anti-immigrant comments (combined effect of Time 4 and 5). In the forum, participants who make their donation decision in Time 4 or 5 in the exposed group have observed 50% or 66% of comments with anti-immigrant views, respectively. The results of these times are combined and compared to the donation frequency of the rest of the times. The comparison shows that the combined donation probability is the same before and after observing half of the comments as anti-immigrant views (Δ=-0.001, $p = 0.98$), and also not different from donation levels in the unexposed group. To look for non-linear effects of the number of comments in the decision to donate, a model with a linear, a quadratic, and a cubic term of the number of comments is constructed (see Table A14). This turns the regression model into a curve. None of the terms has a significant effect on donations. The results are similar for donations to Pro Asyl. The increasing number of anti-immigrant comments does not have a significant effect on the decision to donate to neither of the organizations.

2.3.2 Gender is a Moderator of the Effect of Observing Norm Transgressions

In this section, we investigate whether the effect of the increasing anti-immigrant comments is contingent on the gender of the participant. The analyses use a sample of 2271 instead of 2238 participants because 48 participants have missing information about their gender. The number of missing observations is larger in Times 1 and 2 (8 versus 3 missing observations), although this difference is not significant, neither in the unexposed group ($p = 0.444$) nor in the exposed group ($p = 0.101$). The analysis in this section focuses on donations to the AfD, to which, on average, women donate less than men ($p = 0.004$).

Following hypothesis 3, we should expect women to exhibit behavior consistent with the anti-hate norm longer than men in the presence of increasing anti-immigrant comments. Figure 2.5 shows the frequency of donations in relation to the total number of anti-immigrant comments observed before making the donation decision. The plot uses only observations from the exposed group to keep the number of observations constant in each group. The results using all the observations are depicted in Figure A4.

Women and men donate at similar rates in the baseline Time 1, where no anti-immigrant comments are shown (Δ 0.055, $p = 0.437$). After the inclusion of the first anti-immigrant comment, women begin to donate less than men at all times. The logit model in Table A9 shows that men display a larger predicted probability of donation after seeing one or more anti-immigrant comments (Pr(Donation)=0.2) when compared to Time 1 (Pr(Donation)=0.145), although the differences are not significant (Δ=-0.059, $p = 0.28$). Women, on the contrary, donate less after the inclusion of anti-immigrant comments than in the baseline Time 1 (Δ=-0.059, $p = 0.28$). The predicted probabilities of donation for each number of anti-immigrant comments observed and the tests of significance are shown in Table 2.4.

The predicted probability of donation to the AfD for women decreases after the introduction of anti-immigrant comments when compared to having observed no anti-immigrant comments (see Table 2.4). Not only do women donate less than men, but they react in opposite ways to the increasing number of xenophobic comments. After the addition of only one anti-immigrant comment, the predicted probability of donation decreases in -0.106 for women ($p = 0.093$), whereas it increases for men. This goes in line with hypothesis 3 that we should expect women

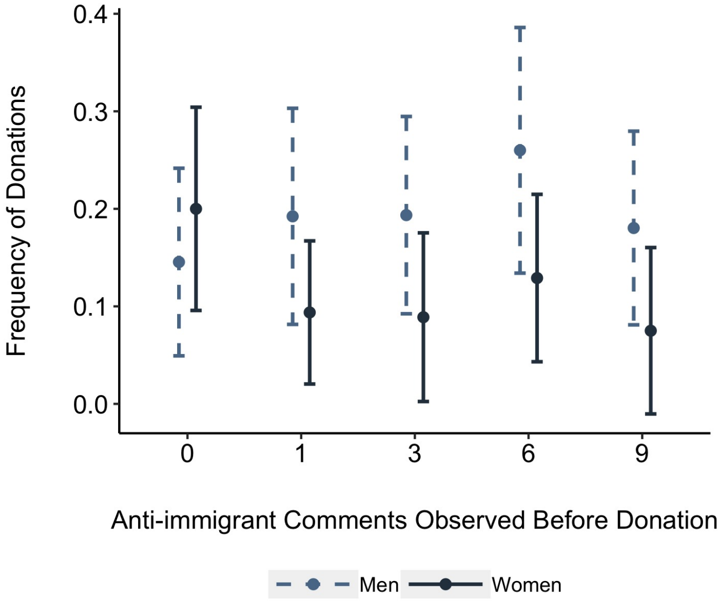


Figure 2.5: Frequency of donation in relation with the total number of racist comments observed before making the donation decision. Error bars at the 95% confidence interval. The dashed line represents the results for men, and the solid line the results for women (Obs=551).

to show more reluctance than men to conform to the trending anti-immigrant norm.

Table A10 and Figure 2.6 show the results from the analysis comparing results in the exposed and exposed group (Group × Times).

On average, women donate less than men in the unexposed group, and this difference is even larger in the exposed group (Δ=-0.071, p = 0.02). This result favors an interaction effect of gender and exposure to anti-immigrant comments in which women react by donating less when anti-immigrant comments are observed and men do the opposite. For women in the unexposed group, the frequency of donations is similar in Times 1 and 3, and increases in Times 4 and 5 (Time 5, Δ=-0.115, p = 0.088). In the exposed group, the predicted probability of donation decreases after the inclusion of one anti-immigrant comment in Time 2 in more than 10% (Δ=-0.106, p = 0.093), as well as in Time 3 (Δ=-0.11, p = 0.096), Time 4 (Δ=-0.07, p = 0.23), and Time 5 (Δ=-0.125, p = 0.060). When compared to Times in the unexposed group, all donation times in the exposed group show a lower predicted probability of donation(with a maximum difference in Time 5, Δ=-0.145, p = 0.03). On the contrary, men display a general tendency to increase their frequency of donations after observing anti-immigrant comments (see Table A11).[14] The predicted probabilities of the interaction effects are shown in Table A12.

[14] The results of the logit models are reported in Table A10. In the logit model, the interaction effect of Time × Group is tested for both gender separately and in a model with all observation, which includes a 3-way interaction term (Times × Group × Gender). Men donate significantly more in Time 4 when exposed to anti-immigrant comments (logit model, $p < 0.05$) than in Time 1. Women donate significantly less in the exposed group in Time 4 ($p < 0.10$), and Time 5 ($p < 0.05$). The 3-way interaction is also significant for Time 4 ($p < 0.01$) and Time 5 ($p < 0.05$) in the exposed group for women. This means the different reactions of women between exposed and unexposed is significantly different from reactions of men.

Table 2.4: Predicted probability of donation by gender and number of anti-immigrant comments observed before donation decision, first differences (to baseline), and second differences.

	Pr (Donation)	Difference to 0 (1st Difference)	2nd Difference
Men 0 Comment(s)	0.145		
	(0.048)		
Men 1 Comment(s)	0.192	0.047	
	(0.054)	(0.072)	
Men 3 Comment(s)	0.193	0.048	
	(0.050)	(0.069)	
Men 6 Comment(s)	0.260	0.115	
	(0.062)	(0.078)	
Men 9 Comment(s)	0.180	0.035	
	(0.049)	(0.068)	
Women 0 Comment(s)	0.200		
	(0.051)		
Women 1 Comment(s)	0.093	-0.106 †	0.153
	(0.036)	(0.063)	(0.096)
Women 3 Comment(s)	0.088	-0.111 †	0.159 †
	(0.042)	(0.067)	(0.096)
Women 6 Comment(s)	0.129	-0.071	0.186 †
	(0.042)	(0.067)	(0.103)
Women 9 Comment(s)	0.075	-0.125 †	0.160 †
	(0.042)	(0.066)	(0.095)

Notes: The first column shows the predicted probability of donations to AfD for each gender depending on how many anti-immigrant comments were observed before the decision. The second column shows the difference in predicted probabilities compared to 0 anti-immigrants comments observed. The third column tests whether this difference is different for men and women (Gender × Number of Anti-immigrant Comments). Standard errors in parentheses. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test.

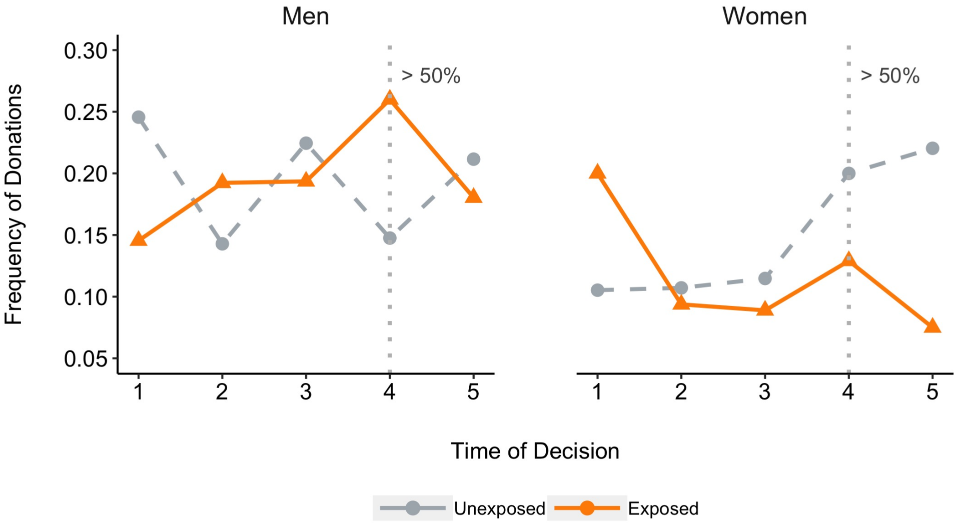


Figure 2.6: Frequency of donations to the AfD in every stage of the forum for men and women. The solid orange line represents the online forum pages where participants are exposed to an increasing number of anti-immigrant comments (exposed group), compared to the unexposed group (dashed grey line). From Time 4, the proportion of anti-immigrant comments is 50% or more, as shown by the dotted line. The number of observations is 1107. Only observations from the AfD condition are used.

Table 2.5: Probability of donation of women by Group (Exposed and Unexposed groups) and Time of donation (Times 1 to 5).

	Pr (Donation)	Difference to T1 (1st Difference)	Difference (Exposed vs. Unexposed)
Unexposed Time 1	0.11		
	(0.041)		
Unexposed Time 2	0.11	0.002	
	(0.041)	(0.058)	
Unexposed Time 3	0.12	0.009	
	(0.041)	(0.058)	
Unexposed Time 4	0.20	0.095	
	(0.054)	(0.068)	
Unexposed Time 5	0.22	$0.115^{\dagger}$	
	(0.054)	(0.068)	
Exposed Time 1	0.200		0.095
	(0.051)		(0.066)
Exposed Time 2	0.09	-0.106 †	-0.013
	(0.036)	(0.063)	(0.055)
Exposed Time 3	0.088	-0.111 †	-0.026
	(0.04)	(0.067)	(0.059)
Exposed Time 4	0.129	-0.071	-0.071
	(0.042)	(0.067)	(0.069)
Exposed Time 5	0.075	-0.125 †	-0.145 *
	(0.042)	(0.066)	(0.068)

Notes: The first column shows the predicted probability of donations to AfD for each gender by group and time of donation. The second column shows the difference in predicted probabilities compared to time 1 for each group. The third column tests whether this difference is different for each group (Group × Time). Standard errors in parentheses. Significance levels: $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test.

Table A15 shows the results of a linear regression model of donations to the AfD with a linear, a quadratic, and a cubic term of the number of anti-immigrant comments. The linear effect shows a significant negative effect of the number of observed anti-immigrant comments (regression model with covariates of demographic characteristics, $p = 0.04$), as well as a positive quadratic regression coefficient ($p = 0.06$). The results of the quadratic coefficient indicate a curvilinear relation between behavior and the observed norm transgressions. This result speaks in favor of a threshold mechanism in the behavior of women, in which donations decrease after the inclusion of the first norm transgressions and increase after a certain threshold.

2.4 Discussion

Over the last decades, a norm against the public expression of views considered politically incorrect developed. As a result, certain views became regarded as socially undesirable. Norms of public speech have always existed; however, anti-prejudice norms became more strictly enforced in recent years and, as noted by Mudde (2004), "most notably with reference to racism" (p. 554). The public expression of certain attitudes became particularly disapproved of in the political arena and is nowadays generally regarded as *Stammtischparolen* or pub talk. After the delegitimization of the use of biological racism in politics since the mid-twentieth century (Rydgren, 2005; Mendelberg, 2001; Pettigrew, 1958; Dovidio & Gaertner, 1986), expressing racist views in public became more and more socially disapproved. A social norm was established.[15]

The anti-hate norm has tangible implications. For example, the adoption of the anti-hate norm could explain why support for political parties with openly racist agendas has seen a declined since the mid-twentieth century when political agendas with a fascist imprint became stigmatized (Ignazi, 1992). Yet, variations of the strength of the norm occur locally and temporarily (Paluck, 2009a, 2009b). The increasing levels of online hate speech and the emergence of new rightwing populist parties are among the many factors that have challenged the anti-hate norm in past years. Nevertheless, the dynamics how change happens have yet to be investigated. This study examines how the probability of adoption of previously counternormative behavior depends on the perception of increasing social acceptability in the local context. In short, it aims to measure the relationship between observing an increasing number of norm violations and the individual decision to also transgress the norm.

We have proposed two different hypotheses of how the dynamics of normative change, a breakdown of the anti-hate norm, might be. Hypothesis 2 states that the behavioral response, i.e., donations, should vary linearly as the number of anti-immigrant comments in the forum increases. The analysis shows a lack of reaction of the participants to the increasing comments in both AfD and Pro Asyl. Participants in the experiment behave very similarly in the baseline of zero anti-immigrant comments, in the unexposed group, and in the forum where participants were exposed to anti-immigrant comments. Likewise, no support was found for a threshold relation between the number of norm transgressions and donations (hypothesis 1). Taking all things into consideration, some aspects of the experimental design might not have been adequate to test the hypotheses. For example, the decision to donate might not be a good proxy for the endorsement of anti-immigrant attitudes. Participants could have perceived the norm as changing, but still do not carry over this perception to the donation decision. Also, the heterogeneity in responses to the increasing anti-immigrant comments might be larger than expected. Yet, the design presented in this work cannot distinguish between possible heterogeneous treatment effects.

[15] As Rydgren (2005) explains, after the Second World War there was a period of de-legitimization of the extreme right and its ideological body in Western Europe. Particularly "biologically based racism" (Rydgren, 2005, p. 413) was marginalized.

The results show support for hypothesis 3, which stated that women will show more reluctance than men to support anti-immigration. We use gender as a proxy for normative sensitivity.[16] When initially exposed to counternormative comments, women react by decreasing their likelihood of donation to the AfD, which indicates a focusing effect of the first norm transgressions (Krupka & Weber, 2009). On average, women donated less to the AfD when exposed to the anti-immigrant comments than in environments where they were not exposed to norm violations. Men, on the other hand, seem to react positively to observing anti-immigrant comments and slightly increase donations, although this effect is inconsistent. Women therefore donate less than men and even less after being exposed to anti-immigrant comments. The results support the idea of a different reaction of men and women when normative concerns are highlighted in the context.

The experiment has two main limitations: i) problems with the operationalization of the *unraveling norm*, and ii) problems with the implementation of the treatments. First, the only factor that varies throughout the different forum pages is the number of comments that are violations of the anti-hate norm. This is how the breakdown of the norm is operationalize in the experiment. However, the effect of the increasing norm violations might depend on many factors that are not taken into account in the design. One factor is the group of reference, that is, people to whom the norm applies. A social norm is always a belief about the expected behavior in a given group of people; therefore, the reference group should be well defined in order for the treatment to work. In the experiment, the online forum is considered a self-contained environment. Nevertheless, if participants did not consider those who committed the norm violations in the forum as part of their reference group, they will not be more prone to conformity, but less.

Second, the implementation of the treatments could have added even more uncertainty about the group of reference. For instance, participants do not know how many others in total participate in the forum. They do not know either how many of the total population of participants are of one type or the other; thus, the sensitivity to the number of hate comments might depend on the participant's prior belief about the distribution of others conforming to the anti-norm. This makes it difficult to detect an average treatment effect because prior beliefs greatly moderate the effect of the experimental intervention. Unfortunately, the study did not gather information about the participants' normative beliefs before or after the experimental treatment.

The results of this study might inform research on how social norms intervention might fail and why. First, results in this study draw attention to the different effects that descriptive norms can have in behavior. Depending on the prior beliefs or sensitivity to the norm, among other factors, observing norm violations might have different effects. They can bring the norm into focus or they can alter the perception of the strength of the norms. Both situations result in different effects; while the former might prompt people to behave in accordance with the norm, the latter has the power to build or unravel the existing social norm. These two effects have to be taken into account when designing social norms interventions, and when analyzing how a

[16] The relationship between gender, normative sensitivity, and motivation to control prejudice are addressed in section 2.1.

situation is regulated by a descriptive social norm. Future research needs to take into account that the reaction to the anti-immigrant comments might have a very high heterogeneity, and it needs to pay particular attention to the distribution of individual thresholds.

The result of a different reaction between men and women to the increasing number of anti-immigrant comments is in the line with the previous empirical observation of women pattern of support to RRPs. In most countries, men are more likely to vote for right-wing populist parties than women (Immerzeel et al., 2015). Different explanations have been suggested to account for this phenomenon: different policy preferences of women or the normative concerns pose for these parties. The *different preferences* assumption does not entirely explain the gender gap, since men and women tend to have similar preferences in crucial matters for right-wing populist votes, such as nativism, authoritarianism, or anti-elitism, and women vote less for these parties even if they agree with their policies (Harteveld et al., 2015). The results in this experiment are in line with the *normative concerns* hypothesis and suggest that women are perhaps more motivated to react to a social norm proscribing public support for political parties with openly xenophobic agendas.

Chapter 3

The Breakdown of Anti-Racist Norms: A Natural Experiment on Normative Uncertainty after Terrorist Attacks[1]

On 18 July 2016, a 17-year-old armed with an axe attacked passengers on board a train heading to Würzburg in the southern part of Germany. Six days later, on 24 July, another attacker injured several people and killed himself when he detonated a backpack bomb in Ansbach, near Nuremberg, in the first Islamist terrorist suicide attack in Germany. Both attacks were later claimed by the Islamic State (IS).[2] In the first attack, five people were seriously wounded. After the first attack, a video released by IS called the attacker "a soldier of the Islamic State". In the second attack, 15 people were injured and the attacker died. The IS-linked Amaq news agency stated that the attacker was an IS soldier (Europol, 2017). The two consecutive terrorist attacks fuelled an already heated public discussion on German policies on migration issues and the European refugee crisis.[3] In the aftermath of the events, media coverage focused on the dangers of opening borders, broadening a public debate on these policies.[4]

After terrorist attacks, hate crimes often follow suit (King & Sutton, 2013; Disha, Cavendish, & King, 2011; Byers & Jones, 2007). The effect is particularly noticeable when the attacker is characterized as a member of a social or religious minority, as exemplified by the wave of anti-Muslim hate crimes that followed the 9/11 terrorist attacks (Disha et al., 2011; Byers &

[1] This chapter is based in a working paper with the same title coauthored with Fabian Winter. The article is currently under review.

[2] The attacks were known, respectively, as the Würzburg train attack (*Anschlag in einer Regionalbahn bei Würzburg*) and the Ansbach bombing (*Sprengstoffanschlag von Ansbach*). The attacks were respectively the 4th and 5th Islamist attacks directed to civil population in Germany, and the ones with the largest number of injured people to the date. The Ansbach attack remained the incident with the most non-fatal injuries until the 19 of December of the same year when a truck was driven into a Christmas market leaving 12 deaths and 56 injuries.

[3] The so-called European refugee crisis is the name given to a period beginning in 2015 when the number of refugees and migrants entering the European Union (EU) dramatically increased.

[4] This is documented in news coverage of the terrorist attacks and German political events during July and August, 2016 (e.g., Connolly, 2016; Oltermann, 2016; Hack, 2016).

Jones, 2007; Hanes & Machin, 2014), the increase in violence against refugees linked to Islamist attacks in Germany (Jäckle & König, 2018), or the escalation of racial and religious hate speech on Twitter after a murder committed by Islamist extremists in the United Kingdom (Williams & Burnap, 2015). More generally, formal and informal norms of "civic behavior" seem to erode after such attacks and behavior that was not acceptable before becomes more frequent in the aftermath.

We will explain the erosion of civic behavior by focusing on the most immediate public reaction to terrorist attacks that can usually be observed in social media: the public expression of prejudice gains traction in online environments (Awan & Zempi, 2017; Burnap et al., 2014). We will refer to this as *hate speech*, which is speech intended to promote hatred on the basis of race, religion, ethnicity, or sexual orientation (Gagliardone et al., 2015). As of now, however, little is known about the mechanisms causing this increase. It is well established through observational studies that terrorist attacks have a profound impact on xenophobic attitudes (Legewie, 2013; Walters et al., 2016; Echebarria-Echabe & Fernández-Guede, 2006; Boomgaarden & de Vreese, 2007). The role of the terrorist attacks as trigger events of xenophobic attitudes has lead many scholars (e.g., Hanes & Machin, 2014; Awan & Zempi, 2017) to *assume* that the rise in online hate results from the change in attitudes. The widespread *attitudinal change argument* states that terrorist attacks increase xenophobic attitudes and anti-immigrant sentiment because people perceive terrorist attacks carried out by out-group members as inter-group threats. This leads to an increase in prejudice (Riek et al., 2006), and, so the argument goes, to an increase in hate speech as a direct consequence of the change in attitudes.

We contribute to this debate by suggesting, and empirically testing, a complementing mechanism. We argue that the attitudinal change argument obviously makes an important contribution, but it misses a crucial point: hate speech is a communicative act and, as such, it is regulated by social norms. Social norms play a decisive role in containing the public expression of prejudice such as xenophobic attitudes. Previous research shows that people veil their true attitudes strategically when they believe their preferences are not socially accepted (Kuran, 1995). They avoid disclosing their political opinions to those whom they believe to hold opposite views (Cowan & Baldassarri, 2018) and express less racist opinions after overhearing someone else doing so (Blanchard et al., 1994). Since there are very few formal rules in many online contexts, social norms play a crucial role in these domains. Recent evidence suggests that emphasizing a social norm against the use of hateful language (Álvarez-Benjumea & Winter, 2018; Cheng et al., 2017) or informally sanctioning it (Munger, 2016) reduces the incidence of online hate speech. People are less likely to express hate when a social norm against its expression is unequivocally in place and changes in expectations about the norm can translate into changes in behavior. The relation between norms and behavior is often direct and the effect of anti-prejudice norms can be observed even when attitudes remain unchanged. For example, Paluck (2009a) found that decreasing social acceptability of prejudice reduced its public expression without affecting personal beliefs. In a field experiment conducted by Blanchard et al. (1994), hearing only one person either condemn or condone racism led subsequent participants to either to endorse or to oppose racist views.[5]

A central building block of the suggested mechanism is an increase in normative uncertainty caused by terrorist attacks. They lead to a situation of confusion in which previous norms are challenged and do not seem to apply any longer. These situations of little normative guidance have been described as a state of *anomie* by Durkheim (1897), but in his conceptualization apply to a much broader state of lawlessness. We will therefore interchangeably use the terms *local anomie* or normative uncertainty, which should be understood to apply to a more narrow, local set of norms. In this state of local anomie after terrorist attacks, people become more receptive to normative cues. They more readily follow the example of the hate speech of others by expressing their own attitudes more blatantly. We thus argue that increases in hate speech after terrorist attacks not only depend on their impact on individual attitudes, but also on the uncertainty of the norm against hate speech.

To test the validity of our mechanism, we report the results of a unique combination of a natural experiment and a lab-in-the-field experiment. This allows us to test an empirically challenging idea: the local anomie created by the attacks is not directly observable, such that both an increase in negative attitudes and an increase in norm uncertainty could result in the same effect on hate speech. We exploit the occurrence of the two consecutive Islamist terrorist attacks in Würzburg and Ansbach in Germany (the natural experiment), combined with a pre- and post-attacks lab-in-the-field online forum in which we exogenously manipulate a descriptive norm against the use of hate speech. The lab-in-the-field experiment ensures proper randomization of participants into experimental conditions, which guarantees the same level of attitude change across conditions and hence isolates the effect of social norms.

The experimental online forum covers discussions on two different social topics: refugees and gender rights including feminism and Lesbian-Gay-Bi-Transgender (LGBT) rights. Participants in the forum are randomized into three different conditions which vary the descriptive norm against hate speech: a mixed, a neutral, and a positive condition. The mixed condition consists of a mix of comments from friendly language to actual transgressions of the anti-hate norm. Because this condition does not signal any specific descriptive norm, it does not reduce the norm uncertainty produced by the terrorist attacks. In the neutral condition, we show only neutral or positive comments to the participants. This biases the perception of how many others use hate speech, thus creating a behavioral regularity that signals the existence of a descriptive norm against hate speech. To emphasize further the descriptive norm, we create a positive condition in which only positive comments about the respective minority group are shown. This design thus allows us to compare the effect of the terrorist attacks in contexts where a descriptive norm against the use of hate speech is clear to contexts in which the norm is uncertain. Randomization also allows us to isolate the effect of social norms from the effect of attitudinal change.

In the next sections, we describe the political context in which the attacks took place, we lay

[5] For a discussion on the relationship between social norms and behavior and how normative perceptions can be more malleable than attitudes, see Tankard and Paluck (2016).

out the theoretical foundations of the proposed mechanism, outline our identification strategy, and present the results of our analysis. Finally, the implications of the findings for the dynamics of online hate after terrorist attacks and the implications for the literature on social norms of communication are discussed.

3.0.1 Political Situation of the Attacks

The Würzburg and the Ansbach terrorist attacks can be contextualized in the European refugee crisis. Unrest in the Middle East, especially in Syria and Iraq, caused a massive displacement of people fleeing war and political instability, pushing large numbers of refugees to the surrounding countries and to Europe. These people included mostly asylum seekers, but the possibility of hostile individuals, including IS militants, also reaching the EU was widely discussed in the media. Across Europe, the mass immigration generated sympathetic responses towards the newcomers, but also precipitated fears related to the capacity of assimilation and fuelled anti-immigration and populist discourses (Andersen et al., 2017). Public opinion linked the crisis to a surge in Islamist terrorism and fed a narrative around immigrants as threatening security and western values, a frame frequently reinforced by the media (Greussing & Boomgaarden, 2017).[6] Negative attitudes toward minorities became more common (Wike, Stokes, & Simmons, 2016), and the widespread attitudes of prejudice increased a fear of an upswing in hate crimes.[7]

As the European country that welcomed the highest number of refugees, Germany's reaction attracted a lot of attention. Chancellor Angela Merkel pursued an open border policy, but many others in the country challenged these policies, and anti-immigration parties gained support.[8] In addition to the political reactions, public acceptance of immigrants in Germany decreased from 2015 to 2016 (Czymara & Schmidt-Catran, 2017) and violence directed towards refugees has been on the rise (European Union Agency for Fundamental Rights, 2016). After the attacks, rising levels of online hate speech and the fear that they might cause physical attacks became a matter of concern (Müller & Schwarz, 2017). The hashtags #Merkelsommer – *Merkel's summer* – and #Merkelmussweg –*Merkel should go* – were among the most-discussed topics on Twitter, asking for the resignation of chancellor Angela Merkel, along with general messages against refugees, such as the Twitter hashtag #refugeesNOTwelcome (Kreis, 2017) gaining popularity.[9] Recurring

[6] The number of suspects arrested for Islamist terrorism has steadily increased from 2012 to 2016 (Europol, 2017). 2016 alone resulted in 13 attacks and 135 people being killed. A predominant explanation in European and German media for attacks akin to the ones described here is the refugees' religious and ethnic background. See, for example, Stürmer, Rohmann, Froehlich, and van der Noll (2018), who analyze the 2016 New Year's Eve sexual assaults in Cologne and show how German mass media connected the attacks against women to refugees and their religious and cultural background.

[7] Police records confirm an increase in hate crimes in the period, with documented cases of discriminatory or hateful acts targeting refugees and immigrants (European Union Agency for Fundamental Rights, 2016).

[8] For instance, the anti-immigration party Alternative for Germany (AfD) was founded in 2013 and narrowly missed the minimum threshold of 5 % to become part of the parliament in that year's federal elections. In the federal election in 2017, the votes for this party rose to almost 13%.

[9] The first tweet that contained #refugeesnotwelcome appeared on 10 August 2015 (Kreis, 2017). The hashtag has been recurrently used since then in relation to the refugee crisis, including variants such as #rapefugeesnotwelcome after the New Year's Eve sexual assaults in Cologne in 2016.

political rallies against the chancellor and her policies were held, and many politicians released comments attacking the open-border policy.

3.1 Terrorist Attacks Trigger Normative Uncertainty

Figure 3.1 illustrates the different causal pathways that explain the link between terrorist attacks and a rise in hate speech. The left diagram (Terrorist Attacks - Attitudes - Hate Speech) correspond to the *attitudinal change argument.* Terrorist attacks increase xenophobic sentiment towards those groups linked to the perpetrators of the attacks and consequently lead to more hate speech against that group. The connection between terrorist attacks and attitudinal change is empirically well corroborated in previous research (e.g., Legewie, 2013; Walters et al., 2016; Echebarria-Echabe & Fernández-Guede, 2006; Boomgaarden & de Vreese, 2007), while the link from attitudinal change to hate speech is usually only assumed and not directly tested (Hanes & Machin, 2014; Awan & Zempi, 2017). Our experiment "controls away" the effect of attitudinal change via randomization into the experimental conditions, which allows us to focus on the alternative mechanism.

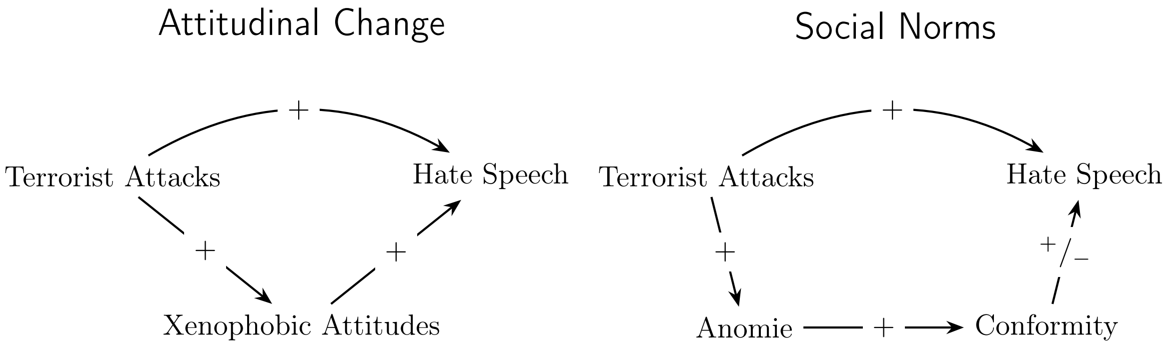


Figure 3.1: Pathways of the *attitudinal change* mechanism (left) and the proposed *social norms* mechanism (right). The curved arrows represent a macro relationship, the straight arrows the micro mechanism.

Our mechanism is sketched in the diagram on the right (Terrorist Attacks - Anomie - Conformity - Hate Speech). Terrorist attacks create normative uncertainty and create a situation of confusion, which is local anomie. The previous consensus about the social norm against the public expression of hate erodes and does not provide guidance for behavior. We therefore find ourselves uncertain about what behavior is expected from us, and become more receptive to normative cues. As a result, social conformity increases. The resulting effect on hate speech would depend on the available cues in the social context. Our mechanism thus relies on several elements: hate speech is regulated by social norms, terrorist attacks increase anomie, and anomie promotes the search for cues, particularly the behavior of others. The different building blocks are explained in this section.

Norms regulate hate speech Social norms are shared informal rules that provide the standard of behavior in a social context (Bicchieri, 2006). Their importance for the public expression of hate has been repeatedly established (e.g., Ivarsflaten et al., 2010; Blinder et al., 2013; Crandall et al., 2002), which makes the expression of prejudice more likely in a private than in a public context (Ford, 2008; Blanchard et al., 1994). The anti-hate norm can be found in the literature under different names, such as egalitarian norm (Crandall et al., 2002), norm of racial tolerance (Weber et al., 2014), anti-racist norm (Ivarsflaten et al., 2010), or anti-prejudice norm (Blinder et al., 2013). The emergence of the norm against hate shapes the way people discuss social issues in public, and also shaped political discourses over the past decades (Mendelberg, 2001).[10] Social norms are inherently ambiguous and individuals look for cues in their environment to asses the social acceptability of a behavior like verbalizing hate. A main source of normative information is what others are doing, e.g., their publicly observable behavior (e.g., Tankard & Paluck, 2016; Bicchieri, 2016; Krupka & Weber, 2009).

Descriptive norms serve as cues for normative behavior Behavioral regularities, known as descriptive norms in the literature, are the most available way of perceiving social norms. In past studies, observing social referents, both in media and in real life (Paluck, 2009a; Paluck, Shepherd, & Aronow, 2016), or simply observing norm-consistent behavior (Blanchard et al., 1994), reduced the expressed prejudice by enhancing the anti-hate norm. Consensus information also changes the perception of the social acceptability of prejudice, leading people to act accordingly. People adjust their reported levels of prejudice after learning the beliefs of the majority (Stangor et al., 2001), and they are more likely to express xenophobic views publicly if they believe others approve of them (Bursztyn et al., 2017). Descriptive norms regulate the decision to use hate speech in online contexts. Users of an online forum are, for instance, less likely to use hateful speech in environments in which a descriptive norm against the use of hate is highlighted (Álvarez-Benjumea & Winter, 2018; Cheng et al., 2017).

Social norms are shared informal rules that provide the standard of behavior in a social context (Bicchieri, 2006). Their importance for the public expression of hate has been repeatedly established (e.g., Ivarsflaten et al., 2010; Blinder et al., 2013; Crandall et al., 2002), which makes the expression of prejudice more likely in a private than in a public context (Ford, 2008; Blanchard et al., 1994). The anti-hate norm can be found in the literature under different names, such as egalitarian norm (Crandall et al., 2002), norm of racial tolerance (Weber et al., 2014), anti-racist norm (Ivarsflaten et al., 2010), or anti-prejudice norm (Blinder et al., 2013). The emergence of the norm against hate shapes the way people discuss social issues in public, and also shaped political discourses over the past decades (Mendelberg, 2001).[11] Social norms are

[10] The dominant norm in the late 1950s in reference to racist speech was one of pro-racist attitudes (Duckitt, 1992). Since then, the public expression of racial prejudice declined over the last decades in western societies (Schuman et al., 1997), and survey respondents were less willing to endorse overt racial prejudices (Huddy & Feldman, 2009).

[11] The dominant norm in the late 1950s in reference to racist speech was one of pro-racist attitudes (Duckitt, 1992). Since then, the public expression of racial prejudice declined over the last decades in western societies (Schuman et al., 1997), and survey respondents were less willing to endorse overt racial prejudices (Huddy &

inherently ambiguous. They are generally not clearly determined so individuals look for cues in their environment to asses the social acceptability of a behavior like verbalizing hate. A main source of normative information is what others are doing, e.g., their publicly observable behavior (e.g., Tankard & Paluck, 2016; Bicchieri, 2016; Krupka & Weber, 2009). .Behavioral regularities, or descriptive norms, often outperform other sources of information (Cialdini, 2007; Horne et al., 2018)

Terrorist attacks create local anomie We argue that terrorist attacks challenge the validity of certain anti-hate speech norms. People question whether it is still required to veil one's anti-migrant attitudes or whether it is now permitted to raise them publicly. In reference to (Durkheim, 1897), we refer to this situation of social disorganization in which preexisting social norms not longer work as local anomie. It can be brought about by events that are disruptive, such as economic crisis, war, or even celebrity suicides (Hoffman & Bearman, 2015). More specifically, events that are perceived as threatening, or are framed as damaging to core values of society, are likely to trigger anomie. The effect increases for heavily publicized events,[12] and obviously terrorist attacks fall in this category.[13] In this situation, individuals seek to regain orientation and thus look at others to form an idea about what is expected.[14] Anomie is not directly observable, but we can, for example, measure the effect of a given event on the perceived strength of the anti-hate norm and how the norm changes in reaction to the event. Events that are exceptional, and therefore disruptive, are likely to cause anomie. For example, after the rather unexpected win of Donald Trump in the 2016 US presidential election, participants in an experiment reported perceiving higher normative acceptability of openly expressing prejudice towards groups targeted during the campaign, but they also adjusted their perception of their own prejudice level by comparing to the new standard (Crandall et al., 2018).

Local anomie increases normative conformity The more uncertain, or anomic, a situation, the more people rely on social cues. Uncertainty facilitates social processes aimed at gaining accurate information about the context, such as self-categorization (Hogg, Sherman, Dierselhuis, Maitner, & Moffitt, 2007), social comparison (Myers, 1982), and, importantly for our mechanism, normative conformity (Deutsch & Gerard, 1955; Willer et al., 2009). Normative conformity means that people pay more attention to their context in an attempt to learn the appropriate behavior. As a result, people are more likely to copy the behavior of those who are believed to

Feldman, 2009).

[12] The level of exposure to media moderates reactions to terrorist attacks, such as increased levels of anxiety (Huddy, Khatib, & Capelos, 2002), in-group favoritism (Traugott et al., 2002), and stereotyping immigrants as a threatening (Boomgaarden & de Vreese, 2007).

[13] In addition to their threatening nature, terrorist attacks are also likely to instill a general sense of uncertainty as to which norms apply by, for example, disclosing previously hidden opinions, or offering a window of opportunity for those individuals who want to frame immigrants as a threat to security to voice their opinion.

[14] Individuals also seek to gain security by reinforcing predominant social categories (Hövermann, Messner, & Zick, 2015), therefore triggering prejudice and xenophobic violence (Jäckle & König, 2018).

represent the majority (Bicchieri, 2016; Willer et al., 2009).[15] In the case of the expression of prejudice, the uncertainty of the norm has been shown to directly affect the degree of conformity. Zitek and Hebl (2007) found that, as the social norm against hate became less clear, participants were more likely to adjust to a confederate's opinion when reporting their own prejudice. A single person can therefore set a normative expectation in situations of normative uncertainty. Furthermore, normative changes in the expression of prejudice are more likely to happen in topics for which the appropriate social norms are unclear (Crandall, Ferguson, & Bahns, 2013). It follows that normative uncertainty does not necessarily imply an increase in the expression of hate, but rather it makes people more open to normative influence. Local anomie can thus either amplify or mitigate the escalation of hate speech after terrorist attacks, depending on the social context.

3.2 Experimental Design: A Combination of a Lab-In-The-Field-Experiment and a Natural Experiment.

Between two waves of data collection in a lab-in-the-field experiment on hate speech in an experimental online discussion forum, two consecutive terrorist attacks took place in Germany. We will occasionally refer to this as the *treatment* in the experimental jargon. We analyze the impact of the terrorist attacks in the hate speech, i.e., speech promoting hatred on the basis of race, religion, ethnicity, or sexual orientation, displayed by our participants. The forum consists of discussions on two different social topics: i.e., gender rights and refugees.[16] The forum experimentally varies the composition of previous comments displayed to the participants, which we will refer to as *experimental conditions.* In the *mixed* condition, which serves as a baseline, no specific norm is signaled. In the *neutral* and *positive* conditions, an anti-hate descriptive norm is highlighted with different strengths. The different experimental conditions offer the opportunity to compare the effects of the attacks in contexts where a descriptive norm against the use of hate speech is highlighted to contexts in which the norm is uncertain. The setting allows us to compare any changes in comments on refugees to changes in comments discussing gender rights to control for common past trends or period effects that could bias the results.

3.2.1 Design of the Experimental Forum

The forum was designed in three steps: i) selection of topics and pictures, ii) collection and classification of initial comments that would later be shown to the participants and iii) construction

[15] There is a large empirical literature supporting this idea that, whenever facing an uncertain situation, people will look for behavioral regularities. See the literature on "herding behavior" or cascades (e.g., Banerjee, 1992; Bikhchandani, Hirshleifer, & Welch, 1992). For a discussion on how behavioral regularities, herding behavior and uncertainty may create the conditions for the emergence of descriptive norms, see Bicchieri (2006, pp. 213-227).

[16] Originally, we selected comments on refugees, LGBT rights and feminism. Feminism and LGBT rights were grouped together in an umbrella category named gender rights to simplify the analysis.

of the different experimental conditions. The different steps for constructing the forum and collecting the data are depicted in Figure 3.2. In the first step, we collected 200 pictures using a list of keywords[17] from different online platforms. Next, we used an online survey in which we asked 90 respondents to decide which topics and pictures would be controversially discussed from the list of 200 pictures. Following that, we selected the top rated pictures and topics in the survey to construct the forum. This forum was available online in a pre-experimental session to collect a first batch of comments on the pictures. A team of three trained raters classified the pool of 840 comments into three categories: neutral, friendly, and hostile. The different experimental conditions consist of a combination of these original comments. By using exactly the same comments before and after the terrorist attack, we avoid the endogeneity problem that often comes with studying peer effects, in which individual and peer behavior are mutually reinforced (Angrist, 2014).

At the beginning of the experiment, participants were given a user name and an avatar. They were told about the experimental nature of the study, but not the actual purpose of the experiment. They were asked to join the conversations and leave comments on the different posts. Once the experiment started, every participant was consecutively presented with the discussions and asked to leave a comment at the bottom of each thread (see Figure 3.3 for a screenshot of the neutral condition). Importantly, we told the participants that the comments should be readable and on the topic, but intentionally avoided stating any expectations about normative concerns. Also, participants could not see what other participants immediately before them had commented, but only the comments we had previously selected to create the different conditions. This ensures that individual observations are independent, and thus increases the internal validity of our results. Each participant was required to leave a comment on each forum page, with a total of eight comments per participant.[18] Each participant gave a comment on each of the eight pictures illustrating the two topics: five pictures on gender rights, and three pictures on refugees. At the end of the experiment, participants received a code that could be exchanged for the payment in an anonymous manner.

3.2.2 Experimental Conditions

Participants in the forum were randomized into three different experimental conditions: a mixed, a neutral, and a positive condition. Each condition consisted of a different mix of comments, from friendly language to actual transgressions of the anti-hate norm, which varied the descriptive norm against hate speech. Table 3.1 shows a summary of the forum content in the different conditions. The *mixed* condition featured a mix of six comments: 2 positive, 2 neutral, and 2 hostile. This configuration did not signal any specific descriptive norm, and therefore did not

[17] The images were obtained from Twitter and Google during March 2016 and we used a set of tags and keywords. Both German and English were used in the search, as both languages are often used on German social media. The following terms and derivatives were used: Sharia, Multiculturalism, Terrorism, Transgender, Gay, Sexism, Discrimination, Refugees, *Aufschrei*, Immigration, Homosexuality, *Einwanderung*, Diversity, Queer, Begging, Atheism, Islamization, Religion, *Tolerist*

[18] Each participant left 7.8 comments on average.

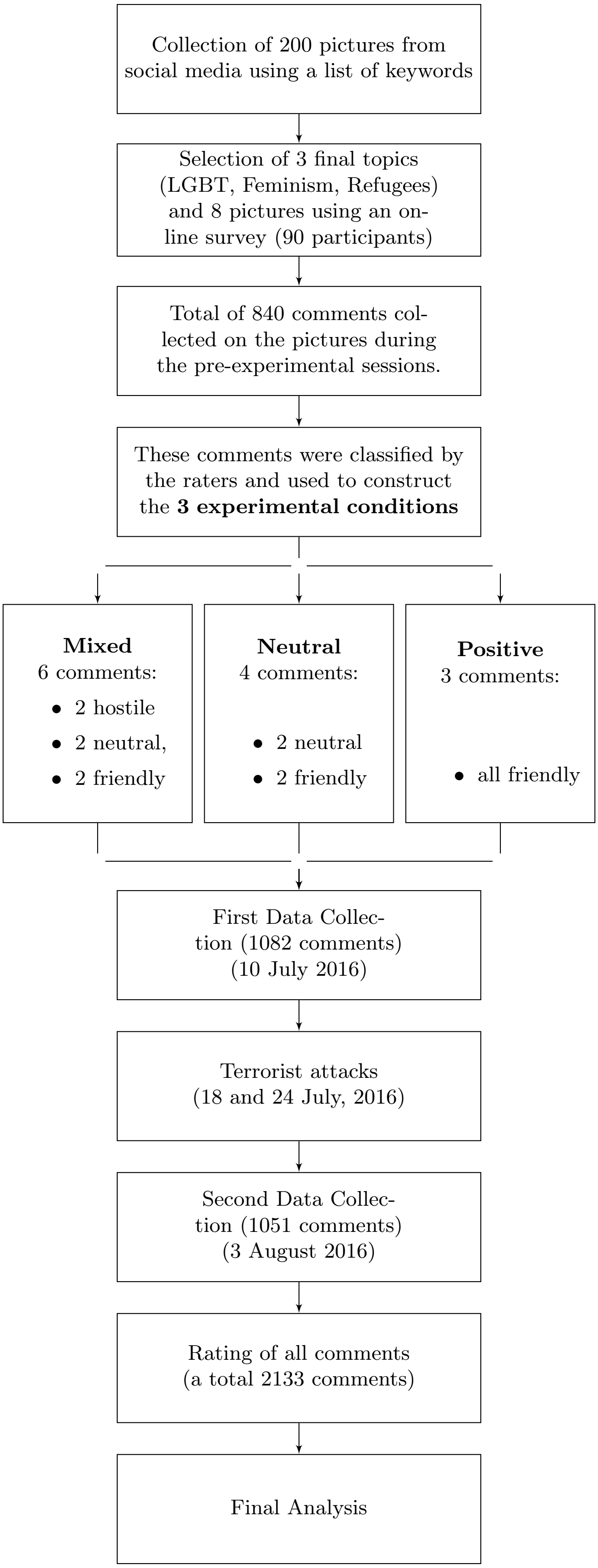
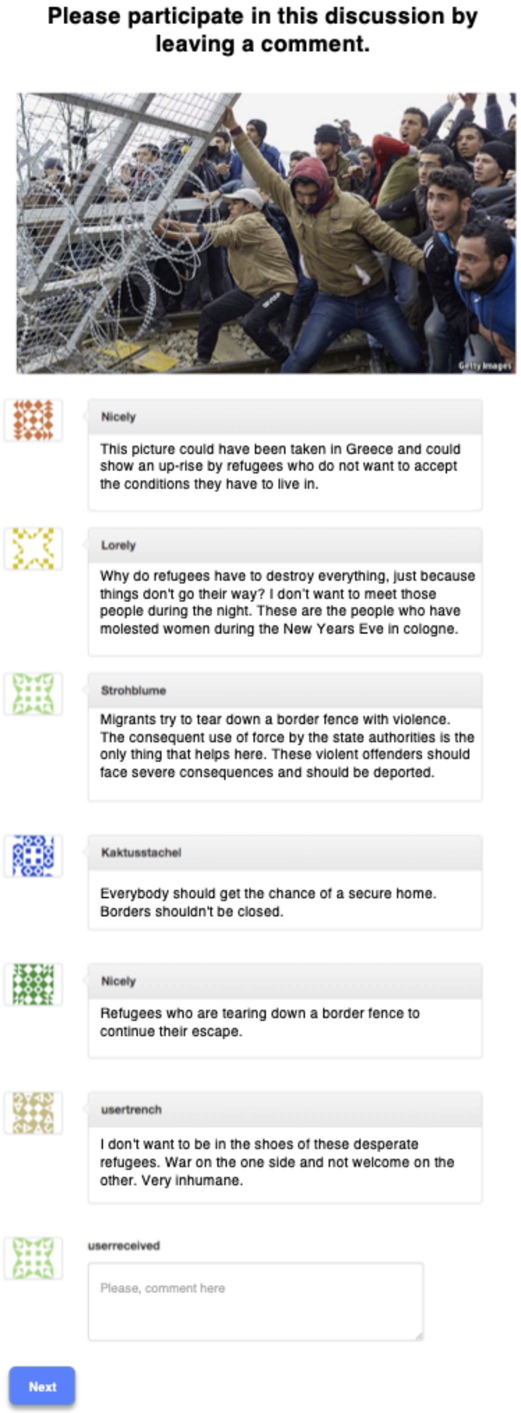

Collection of 200 pictures from social media using a list of keywords
Selection of 3 final topics (LGBT, Feminism, Refugees) and 8 pictures using an online survey (90 participants)
Total of 840 comments collected on the pictures during the pre-experimental sessions.
These comments were classified by the raters and used to construct the **3 experimental conditions**
Mixed 6 comments: • 2 hostile • 2 neutral, • 2 friendly
Neutral 4 comments: • 2 neutral • 2 friendly
Positive 3 comments: • all friendly
First Data Collection (1082 comments) (10 July 2016)
Terrorist attacks (18 and 24 July, 2016)
Second Data Collection (1051 comments) (3 August 2016)
Rating of all comments (a total 2133 comments)
Final Analysis

Figure 3.2: Flowchart of experimental design and online forum set-up.



Please participate in this discussion by leaving a comment.

Nicely

This picture could have been taken in Greece and could show an up-rise by refugees who do not want to accept the conditions they have to live in.

Lorely

Why do refugees have to destroy everything, just because things don't go their way? I don't want to meet those people during the night. These are the people who have molested women during the New Years Eve in cologne.

Strohblume

Migrants try to tear down a border fence with violence. The consequent use of force by the state authorities is the only thing that helps here. These violent offenders should face severe consequences and should be deported.

Kaktusstachel

Everybody should get the chance of a secure home. Borders shouldn't be closed.

Nicely

Refugees who are tearing down a border fence to continue their escape.

usertrench

I don't want to be in the shoes of these desperate refugees. War on the one side and not welcome on the other. Very inhumane.

userreceived

Figure 3.3: Screenshot of the mixed experimental condition (own translation. In German in the original.)

reduce norm uncertainty. In the *neutral* condition, we showed only four comments: two neutral and two positive comments. This biased the perception of how many others use hate speech and created a behavioral regularity that signals the existence of a descriptive norm against the use of hate in the online forum. The *positive* condition further emphasized the descriptive norm by showing only three positive comments.

Condition	Content
Mixed	6 comments: 2 friendly, 2 neutral, and 2 hostile
Neutral	4 comments: 2 friendly and 2 neutral
Positive	3 comments: all friendly

Table 3.1: Summary of the content of the online forum in the different experimental conditions.

Table 3.2 shows the number of comments collected by time of data collection (e.g., before or after the terrorist attacks), experimental condition, and topic.

	Before the Attacks		After the Attacks	
Condition	Refugees	Gender Rights	Refugees	Gender Rights
Mixed	135	227	135	228
Neutral	136	225	135	226
Positive	134	225	123	204
Total Comments	405	677	393	658

Table 3.2: Number of comments (Total=2133) per time of data collection (before and after the terrorist attacks), experimental condition, and topic.

3.3 Research Questions

With the theoretical argument laid out and a thorough description of our experimental methods, we will now present specific research questions. First, we will test whether the terrorist attacks increase the expression of hate speech when no specific norm against hate speech is emphasized. To answer this question we will use data from the *mixed* experimental condition because this condition does not highlight any descriptive norm in particular.

Research Question 1 *Did the terrorist attacks increase the overall level of hate speech in comments about refugees when no specific norm against hate speech is highlighted?*

Second, we test whether the response to the attacks is similar in contexts in which a descriptive norm against the use of hate is highlighted. As discussed above, our experimental approach makes sure that attitudes are randomly distributed and thus on average constant across the experimental conditions. Any remaining differences can thus be attributed to interaction between the anomie created by the terrorist attacks and the normative cues provided by the experimental conditions.

If the terrorist attacks generate normative uncertainty, then this uncertainty should be reduced in the experimental conditions with a strong descriptive norm compared to situations where the anti-hate norm is vague (*mixed* condition). Any differences in behavioral patterns among the conditions can be attributed to the anomie mechanism described above.

Research Question 2 *Did the terrorist attacks increase the overall level of hate speech in comments about refugees when a descriptive norm against hate speech is highlighted?*

3.4 Description of Data, Main Variables and Statistical Approach

3.4.1 Sample

A total of 139 different participants before the terrorist attacks and 135 after the terrorist attacks for our experiment, as well as 577 raters of the comments, were recruited via a crowdsourcing internet marketplace.[19] The experiment was conducted entirely in German and the sample was therefore restricted to residents in Germany. Participants voluntarily registered for the experiment via the platform, which helps us to guarantee that every participant participated in only one condition, either before or after the attack. Unfortunately, the platform has a rather restrictive personal data policy. This lead us to decide not to collect additional data on participants' attitudes, political preferences or individual demographic characteristics. Since we randomize participants into experimental conditions, it shouldn't invalidate our conclusions, but it prevents us from doing any sub-group analyses. We anyways present demographic information on the general characteristics of the workforce for reference in Table A2, which was kindly provided by the platform.

3.4.2 Construction of the Mean Hate Score

We define hate speech as speech intended to promote hatred on the basis of race, religion, ethnicity, or sexual orientation (Gagliardone et al., 2015), i.e., hostile expression of prejudice towards minority groups. The outcome of interest is the level of hostility displayed by the participants, which we refer to as the *mean hate score*, and its changes before and after the terrorist attacks, and across the different conditions.

To construct a measure of hate speech, we asked external raters to rate a set of about 30 randomly chosen comments each.[20]. The raters were given the comments in a randomized order and where

[19] We used the platform www.clickworker.com. It is similar to Amazon MTurk, but with a substatially bigger work force in Germany.

[20] 498 raters rated 30 comments (86.31%) and 79 rated less than 30 comments. The comments received a median of 7 ratings. Only 36 comments received 2 or fewer ratings and only 22 more than 12, thirteen being the maximum number of ratings a given comment received. The inclusion or exclusion of raters that rated less than 30 comments yields no statistically significant changes in the score, as well as the exclusion of comments with extremely low (< 2) or extremely high (> 12) number of ratings.

ignorant of the experimental conditions. Just as the main experiment, the rating task was also completed online using a form provided by us.

Every page of the online form displayed a picture and one comment relating to that picture (see section C.2). We asked the raters to rate the comment on the following scale: *Is the comment friendly or hostile towards the group represented in the picture? (Give a number from 1 to 9 where 1 means very friendly and 9 means very hostile).* Comments with lower scores, e.g., 1 to 4, are therefore affable, with a cordial language, and often express a positive opinion. On the other side of the spectrum, high scores such as 8 or 9 generally imply abusive language, e.g., "I cannot stand gay people. They should have a psychiatric exam" , or the use of hate terms,[21] e.g., "They can continue walking away from Europe. They are not just war refugees, 90 per cent are nothing but **social parasites** who do whatever they want here" (emphasis added). section C.1 contains examples of comments and their classification.

The continuous *mean hate score* for each comment allows us to study subtle variations, as well as changes in the distribution of the score, and serves as the main variable of interest in our study. To measure the level of agreement of the raters, we used the intraclass correlation (ICC), which yields a value of 0.57. We also computed the Spearman Rank Correlation for random subsamples of the ratings with similar results (see section C.2). Table 3.3 gives an overview about the descriptive statistics of the mean hate score.

[21] We defined hate terms as unambiguously pejorative or derogatory expressions. For a comprehensive compendium of hate terms, see `http://hatebase.org` (last accessed on 7 February 2019).

Table 3.3: Descriptive Statistics of the Mean Hate Score

Experimental Condition	Topic	Time	Mean	Variance	Median	Max.	Min.	1st Quant	3rd Quant	IQR
Mixed	Refugees	Before Attacks	3.90	2.10	3.86	7.57	1.25	2.71	4.87	2.15
Mixed	Refugees	After Attacks	4.46	2.62	4.33	8.12	1.43	3.21	5.54	2.32
Mixed	Gender Rights	Before Attacks	3.12	2.26	2.80	8.60	1.00	2.00	4.06	2.06
Mixed	Gender Rights	After Attacks	3.33	2.87	3.00	8.25	1.00	2.00	4.38	2.38
Neutral	Refugees	Before Attacks	3.96	2.28	3.65	8.60	1.25	2.86	5.00	2.14
Neutral	Refugees	After Attacks	3.98	1.76	3.75	7.43	1.33	3.00	4.86	1.86
Neutral	Gender Rights	Before Attacks	2.97	1.82	2.67	7.00	1.00	1.86	3.86	2.00
Neutral	Gender Rights	After Attacks	3.09	2.02	2.73	7.33	1.00	2.00	4.00	2.00
Positive	Refugees	Before Attacks	3.95	3.12	3.62	8.75	1.00	2.50	5.00	2.50
Positive	Refugees	After Attacks	3.84	2.37	3.50	7.56	1.20	2.65	4.73	2.09
Positive	Gender Rights	Before Attacks	2.85	2.55	2.33	8.88	1.00	1.75	3.44	1.69
Positive	Gender Rights	After Attacks	3.04	2.52	2.43	7.80	1.00	1.89	3.76	1.87

3.4.3 Statistical Approach

We will analyze the data from our experiment in three steps. First, we measure changes in the level of online hate speech against refugees before and after the attacks and compare it to changes in comments discussing different topics. This allows us to investigate the direct impact of the terrorist attacks on hate speech against refugees. Second, hate speech is compared across the different experimental conditions to test the idea that the reaction to the terrorist attacks depends on the perceived normative uncertainty. Finally, we will show that this effect is most pronounced for the most hateful comments in our sample.

As described above, each participants gave one comment per picture, and each topic has 2-3 pictures. A comment, i.e., the lowest level observation, is thus nested in participants and topics at the higher level. To accommodate for this nested structure, we estimate the change in the mean hate score in a series of random intercept multilevel linear regression models with two crossed random effects for participants and pictures. We present the analyses in several steps. We first investigate whether the terrorist attacks increased hate speech in comments about refugees when the norm is uncertain (Research question 1). To this end, we analyze only the comments made in the mixed condition. The effect is identified by comparing the level of hate speech before and after the attacks only for comments on refugees, and the interaction effect of the attacks on comments discussing refugees versus the rest of topics (After Attacks × Refugees, see Equation C.1 in the Appendix). To investigate our second research question, i.e., whether people are more receptive towards normative cues in times of anomie, we compare the responses in the mixed condition to the responses in the remaining conditions (see Equation C.2 in the Appendix). Third, we establish that this only holds if the norms are really challenged. Since the anti-hate speech norms in comments about gender rights should not be affected by the attacks, we should observe a significant difference-in-difference in the pre- and post-attacks effectiveness of signaling the anti-hate norm for comments on Refugees and Gender rights. Again, we follow a similar strategy and compare the effect of the neutral and positive conditions before and after the terrorist attacks for comments about refugees, and compare these changes to the changes in the Gender rights category (After Attacks × Refugees × Experimental Condition, see Equation C.3 in the Appendix).

It is important to note that the difference-in-differences is a robust estimator of the effect of the attacks, provided there are no spillovers between the treated and the comparison group. However, previous research shows that social norms are usually susceptible to spillover effects (e.g., Keizer et al., 2008; Reno, Cialdini, & Kallgren, 1993). Accordingly, people exposed to social acceptability of hate speech in one topic could carry this effect over to the rest of the topics. If the spillover effect exists, then the estimated effect of the terrorist attacks on online hate is a conservative estimate.

In the final step, we estimate the differences in the distribution of the mean hate score after the attacks in the different experimental conditions. We do this analysis only for comments discussing refugees. We do this for two reasons. First, to give a more comprehensive picture of the effect of terrorist attacks on hate speech. Second, dividing the hate score into different parts gives us the opportunity to look at changes in the most hateful comments. For each topic and experimental condition, we measure the impact of the terrorist attacks along the conditional distribution of the mean hate score using a quantile regression model (Koenker & Bassett Jr, 1978; Koenker & Hallock, 2001).[22] We report results from the 10th to the 95th percentile. Each quantile estimator, in a manner similar to linear regression, minimizes the sum of residuals (Koenker, 2017)(for the specification of the models, see Equation C.4 in the Appendix). Rather than predicting the mean change of the hate score, it looks at changes at the quantiles of the score before and after the terrorist attacks.

3.4.4 Measures to Ensure External Validity

We take several measures in order to ensure the external validity of the experiment. First, our setting allows us to identify the causal effect of the terrorist attacks on hate speech under the assumption that no other potential events could affect the outcome. To our knowledge, there were no terrorist attacks in Germany in the three months leading to the events under study. Finally, we need participants to have been aware of the events, since the reaction to terrorist attacks has been found to be moderated by news exposure (Boomgaarden & de Vreese, 2007). While we do not have direct evidence, we can use Google searches as an indirect indicator of media attention, and the link between the terrorist attacks and the refugee crisis. Figure 3.4 shows the search interest relative to the highest point on the chart for Germany between 1 June 2016 and 31 August 2016. The search interest should be read as follows: a value of 100 is the peak of popularity for a given search term; a value of 50 means that the term is half as popular; likewise, a score of 0 means the term was less than 1% as popular as the peak. After the terrorist attacks, internet searches in Germany with the keywords *refugee(s)*, *asylum seeker(s)*, and derived terms skyrocketed. Searches that refer to gender rights, namely transgender, feminism, or LGBT, did not show any changes within the period (see Figure A9 in the Appendix).

The terrorist attacks and the second instance of data collection are separated by eight days. The first wave of data collection was on the 10 July 2016, nine days before the first terrorist attack. The second wave of data collection started on the 3 August 2016, ten days after the second terrorist attack. While it is difficult to determine the temporal duration of the response to events on theoretical grounds alone, previous findings from survey-based panel studies suggest that the effect of events is long-lived, and persist over an extended period lasting several months (Lecheler & De Vreese, 2011).

Second, our sample of online workers could rise concerns over the participants' motivation to "do a good job in the eyes of their employer", which would lead to an over-estimation of the effect size. As we have described in the experimental procedure, we took several measures in order to address this limitation: i) participants remained anonymous and no personal data were

[22] Quantile regression assumes that "the distribution of the response can be arbitrarily different" (Koenker, 2017, p. 158) conditional in the treatment variable (i.e., the terrorist attacks).

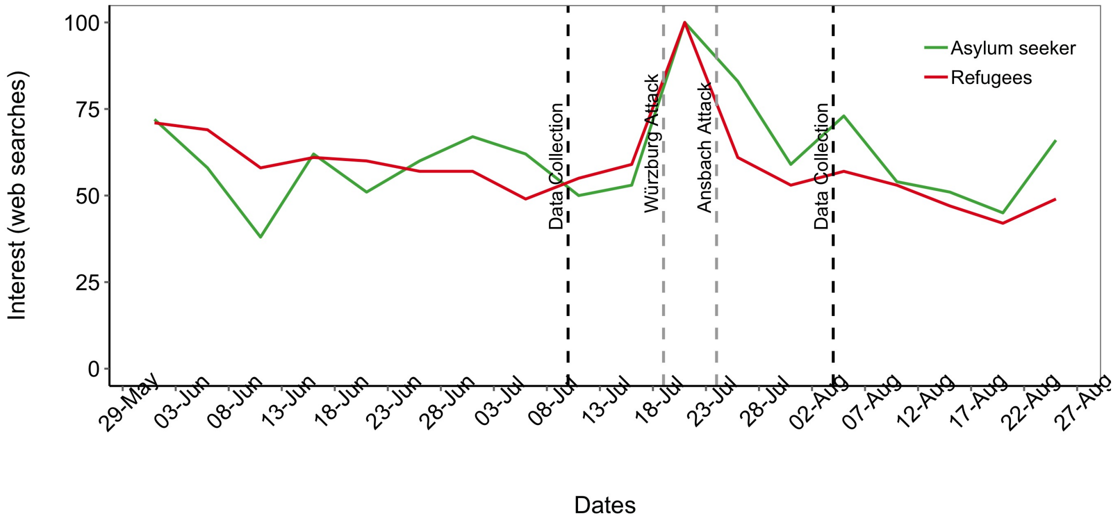


Figure 3.4: Relative interest (web searches) in Germany for a three months period including the data collection times and the terrorist attacks. The plot shows the popularity of the terms from approximately one month before the first data collection time to a month after the second data collection time. The numbers represent the search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. The dashed lines represent the events of interest: dates of data collection and attacks.

collected, ii) participation was voluntary, and iii) payment did not depend on performance. We are confident that these design choices created enough detachment between the forum and the marketplace, and therefore we have no reason to assume that the particular treatment effects are qualitatively changed by our sampling strategy.

Third, the use of online convenience samples could raise concerns, even though it is now commonplace in the experimental social sciences. However, it has been shown that online panels are problematic only to the extent that treatment effects (i.e., the reaction to the terrorist attacks) differ between the online sample and the population of interest (Coppock et al., 2018). Since we do not expect our online sample to perceive the environments any differently than the general population, we are reasonably satisfied that there is a sufficient overlap to allow an interpretation of the results of the field experiment. Our identification strategy would also run into problems if the self-selection into the online sample depended on the treatment. For instance, it would be problematic if workers who have more negative attitudes towards refugees would register for a job on the platform with a higher likelihood after the attacks. While we cannot definitely rule this out, it is difficult to think of a reason of why this should be the case.

It is important to note that these concerns are not limited to online workers. Other sampling strategies, such as university students, could also lead to an increased social desirability bias. Using observational data would potentially solve this problem, but the endogeneity associated with it jeopardizes the proper identification of the treatment effect.

3.5 Results

3.5.1 Hate Speech towards Refugees, but not towards Gender Rights, Increases after the Terrorist Attacks

First, we analyze the effect of the terrorist attacks in the mixed condition (N=725 comments). As described before, the mixed condition features a mix of comments with different levels of hostility and therefore does not signal any specific descriptive norm. This condition does not reduce the anomie produced by the terrorist attacks. We expect hateful comments to increase after the attacks in comments in the category Refugees as a result of the increased local anomie in the aftermath of the attacks. Since one is likely experiencing a wide range of reactions in the "real world", the comments in these condition may be the closest to a natural environment.

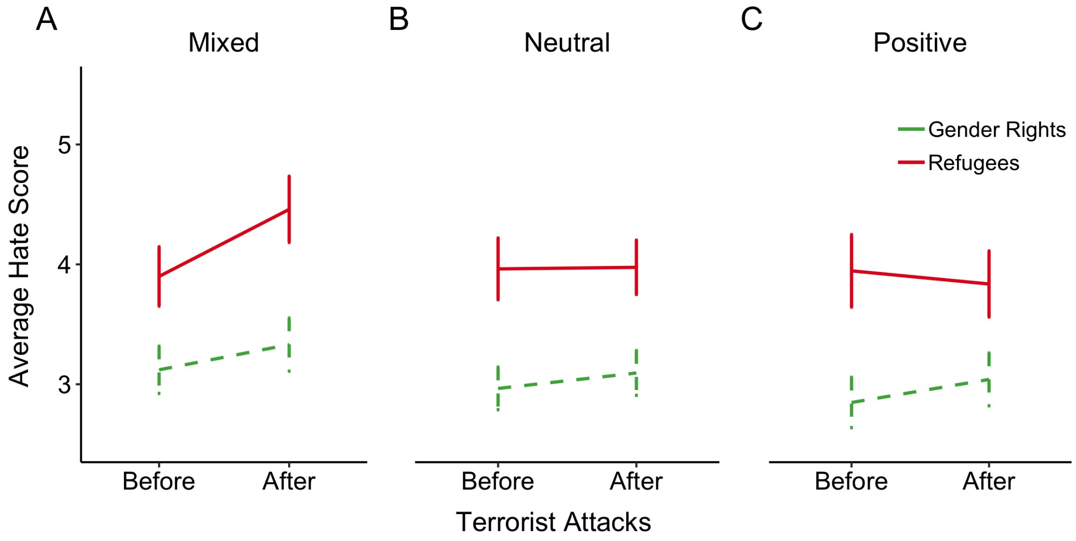


Figure 3.5: Mean hate score by topic before and after the terrorist attacks. From left to right: Mixed Condition (A), Neutral Condition (B), and Positive Condition (C). The solid red line shows the mean hate score for Refugees, the dashed green line for Gender Rights.

Figure 3.5A shows an increase in the mean hate score in comments discussing refugees after the terrorist attacks. Model 1 in Table 3.4 estimates this conditional treatment effect of the terrorist attacks on comments about refugees. The mean hate score increases by 0.56 points following the attacks (p = 0.004). To give an intuition about what these coefficients mean, we can look at how changes in the score are translated into changes in hostility in the comments. A change in one score point can be very noticeable when comparing two comments from a thread on the same topic.[23] A comment with a score of 6 reads: "Get rid of the **funny get-up** ...We are in Europe, or more precisely Germany. Whoever wants to live here as to adapt. Multiculturalism, sure, but not that much." A comment in the same picture, but with a score of 7 reads: "That's a very special bird, a **black barn owl** ... " (in reference to a women wearing a burqa).

[23] Both comments refer to a picture of a woman wearing a burqa sitting in the public transport next to a woman wearing western style clothing. The original comments are in German and translation is our own.

Table 3.4: Main results: Regression estimates and difference-in-difference (DiD) estimates of the effect of the terrorist attacks

	Model 1	Model 2	Model 3	Model 4
Dependent variable: hate score	(Refugees)	(All topics)	(Refugees)	(All topics)
Constant	3.90 (0.19)**	3.12 (0.20)**	3.90 (0.23)**	3.12 (0.20)**
After Attacks	0.56 (0.24)*	0.23 (0.22)	0.56 (0.25)*	0.23 (0.21)
Refugees		0.77 (0.24)**		0.77 (0.25)**
After Attacks × Refugees		0.36 (0.20)†		0.36 (0.19)†
Neutral			0.06 (0.25)	-0.15 (0.21)
Positive			0.05 (0.25)	-0.25 (0.21)
After Attacks × Neutral			-0.55 (0.35)	-0.10 (0.30)
After Attacks × Positive			-0.68 (0.35)†	-0.05 (0.30)
Refugees × Neutral				0.22 (0.19)
Refugees × Positive				0.32 (0.19)†
Refugees × After Attacks x Neutral				-0.48 (0.27)†
Refugees × After Attacks x Positive				-0.67 (0.28)*
Number Comments	270	725	798	2133
Number Participants	90	94	268	274
Number Images	3	8	3	8
$Pseudo - R^2$	.586	.467	.606	.469

Notes: Linear mixed models fit by Maximum Likelihood (ML) with two random effects: participant and image. Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for models with mean hate speech score as the dependent variable. Models 1 and 2 include only comments from the mixed condition. Models 3 and 4 include the full sample. Model 1 shows main effects of terrorist attacks for comments in Refugees. Model 2 shows the effect of terrorist attacks for each experimental condition in comments in Refugees. Model 4 shows the DiD estimate for the effect of terrorist attack in each experimental condition and topic. The table lists mean regression coefficient estimates with standard errors in parentheses and p-values calculated based on Satterthwaite's approximations. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test.

This increase is absent in the category Gender Rights. Model 2 estimates the difference-in-difference between comments on Refugees and Gender Rights before and after the attacks. Before the attacks, comments towards refugees where 0.77 points more negative than those in Gender Rights ($p = 0.029$). After the attacks, comments on Gender rights became 0.23 points more hostile, but this main effect is not statistically significant ($p = 0.28$) and substantially smaller than the 0.59 points increase for post-attacks comments about Refugees. Although the difference-in-difference of 0.36 points is only marginally significant ($p = 0.063$, see the interaction term After Attacks × Refugees in Table 3.4), we believe that this analysis supports the finding that the attacks increase the level of hostility against refugees. Also, because this increase is more pronounced in comments discussing refugees when compared to changes in comments discussing gender rights, the treatment effect of the terrorist attacks seems to be different from a mere period effect. To answer our first research question, the terrorist attacks seem to have increased the overall level of hate speech towards refugees when no specific norm was highlighted.

3.5.2 Descriptive Norms Resolve Anomie After Terrorist Attacks

The central idea of this section is to provide an indirect test of whether the increase in hate speech observed in the mixed environment is caused by the anomie after the terrorist attacks. In this situation, individuals look for normative cues in their environment, such as the behavior of others, to form an idea of what is a socially acceptable response to a situation. In our case, this would translate into a larger effect of the highlighted descriptive norm against hate after the terrorist attacks in comments on Refugees. If this is true, we should expect the increase in hate found in the mixed condition to be reduced or non-existent in the conditions where the norm was highlighted. We should also expect a larger effect of the descriptive norm in comments on Refugees after the attacks when compared to comments on Gender Rights. If, on the contrary, there is no effect of the terrorist attacks on normative uncertainty, we would expect hostility in the forum to increase similarly across all conditions and topics.

To test our idea, we compare the effect of the terrorist attacks on the mixed forum with their effect on the neutral and positive forums for the different topics. Plots B and C in Figure 3.5 depict the average score before and after the attacks in the neutral forum (B) and the positive forum (C). There is no visual pre- and post-attacks difference in hate in neither of these conditions in comments about refugees. As before, we construct a multilevel model with a term for the effect of the attacks in the two censored environments compared to the effect of the attacks in the mixed condition. Model 3 in Table 3.4 again uses only the comments on Refugees and shows that the post-attacks increase of 0.56 points in the mixed condition is offset in both the neutral condition ($\beta = -0.55$, $p = 0.11$) and the positive condition ($\beta = -0.68$, $p = 0.054$). If the participants are confronted with neutral and positive comments only, the stated opinions do not differ before and after the attack ($\beta = 0.014$, $p = 0.95$).[24] This means that the emphasized descriptive norm

[24] The post-attacks mean hate score when participants are confronted only with positive comments is actually

prevented participants from expressing more hateful opinions after the attacks in the neutral and positive conditions. If we analyze only comments made after the attacks, the mean hate score is significantly smaller in the neutral ($\beta = -0.48$, $p = 0.04$) and the positive condition ($\beta = -0.62$, $p = 0.01$) compared to the mixed condition, which means that the main effects of the experimental conditions became significant after the terrorist attacks. These findings are consistent with our theoretical claim that the increase in hate speech only occurs under anomie and when the descriptive norms are ambiguous.

In Model 4, we added a three-way interaction term (After Attacks × Refugees × Experimental conditions) that captures the differential effect of the terrorist attacks in the different conditions by topic. Just as Model 2, Model 4 uses the full sample of comments. Comments towards refugees, in the neutral condition after the attacks, are -0.48 points less hostile ($p = 0.08$), and -0.67 points less hostile in the positive condition ($p = 0.016$) compared to the effect of the experimental conditions after the attacks in the comments discussing gender rights. We thus find (suggestive) evidence for a significantly larger effect of descriptive norms after the attacks in comments on Refugees compared to Gender Rights. These results show that the increase in hate speech in the forum after the terrorist attacks cannot be solely attributed to an increase in negative attitudes towards refugees. An increase in hate resulting from an increase in attitudes would have been consistent across all conditions.

In order to answer our second question, i.e., whether the terrorist attacks increase the overall level of hate speech when a descriptive norm against its use was highlighted, we have looked at the effect of the terrorist attacks in the conditions where the norm was highlighted. In contrast to the post-attacks increase in mean hate score found in the mixed condition, the level of hate speech against Refugees did not change before and after the attacks when the descriptive norm was highlighted. We find that the level of hate, as measured by the mean hate score, did not increase in either of these conditions following the attacks. Furthermore, we find evidence that the effect of the descriptive norms is larger after the attacks when compared to after attacks changes in comments in Gender Rights. We thus conclude that the terrorist attacks did not increase the overall level of hate speech against refugees when a descriptive norm against hate speech was highlighted.

3.5.3 Descriptive Norms Have the Greatest Effect on Extreme Comments After the Terrorist Attacks

Finally, we show that this increase is driven by a shift of the most extreme comments, i.e., already hostile comments with a rating of 7 and above, and not by moderate comments in the range of 3-5 becoming slightly more negative. If this is true, we should expect the largest treatment effect at the higher quantiles of the distribution because they represent the most hateful comments, which can be regarded as violations of the anti-hate norm. We estimate a quantile regression

0.12 (the average hate score is 3.78) points lower than in the pre- attacks mixed condition mean hate score (the average hate score is 4.46), although this is not significant, $p = 0.63$).

model.[25] The quantile regression shows how the magnitude of the effect of the terrorist attacks varies across the different percentiles of this distribution. Figure 3.6 depicts the coefficients of the treatment effect of the attacks on comments about Refugees for the quantiles 0.10 to 0.95 of the distribution of the hate score and the corresponding 95% confidence intervals. The results are shown for all three experimental conditions: mixed, neutral, and positive. Each coefficient corresponds to the change in the τth quantile after the terrorist attacks compared to before the attacks.

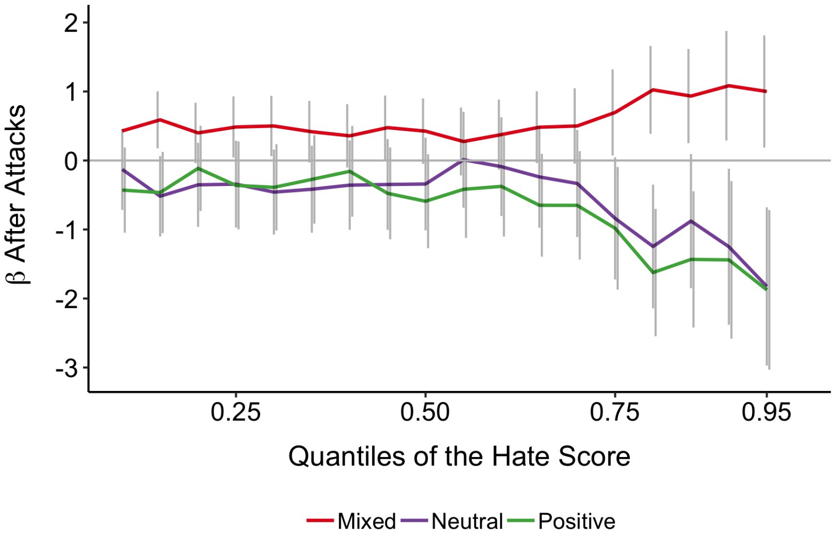


Figure 3.6: The plots depict the estimated model coefficients of the effect on terrorist attacks for quantiles 0.10 to 0.95 in comments on Refugees for all levels of the descriptive norm: Mixed, Neutral, and Positive. The grey vertical lines represent the confidence interval of the quantile regression coefficients for the effect of terrorist attacks with 95% confidence level.

In all three experimental conditions, the effect of the attacks is more pronounced in the highest quantiles of the distribution of the hate score, which suggests that the average increase after the attacks results from extremely hateful comments, i.e., violations of the anti-hate norm becoming more likely in the mixed condition and less likely in the neutral and positive condition. In the mixed condition, the terrorist attacks increased hate speech for all quantiles of the distribution, but this effect is stronger in the most "hostile" quantiles of the distribution. Compared to an average change of 0.56 points in the score in the mixed condition after the terrorist attacks, the 80% and the 90% percentiles have an increase of 1.023 and 1.083 points, respectively. This effect is both statistically significantly different from 0, an from the other conditions.[26] A similar pattern is observed for the mixed and positive conditions: here, the rating of comments in the 75%

[25] We also estimate a linear quantile mixed regression model (Geraci & Bottai, 2014) with participant-specific random intercepts included in the model to account for within-subject dependence. The results using this model are qualitatively similar to the results from the quantile regression we utilize in this paper. The computation of the models with the random term, however, showed several problems of convergence (e.g., converging only at times) due to the small sample size. Therefore the results with the random intercept are not reliable and we decided not to use them.

[26] Figure A10 shows the results of the quantile regression for each combination of topic and experimental condition, and the corresponding density plots.

quantile and above is at least 1.50 points less negative towards refugees after the attack.

3.6 Conclusion

The main theoretical argument of our study claims that extreme events, such as terrorist attacks, create normative uncertainty or anomie about which acts are socially permissible. This uncertainty leads people to search for cues in their environment on how to behave, and the behavior of others provides such cues in the form of descriptive social norms. Depending on these social cues, the prevalence of norm violations may increase, stay the same, or even decrease. Our theory thus provides a complementing mechanism to the widely accepted attitudinal change argument about the reactions to dramatic events. We now believe that attitudinal change only materializes in public transgressions of social norms if the norms are challenged by the event.

We apply our reasoning to explain the erosion of norms of civic conversations after terrorist attacks in an online context. A number of studies show the link between terrorist attacks and an increased level of online hate speech (e.g., Awan & Zempi, 2017; Burnap et al., 2014). Our study empirically confirms this finding. We extend the current state of research, which has established a link between terrorist attacks and increased negative attitudes towards the attacker's social group in empirical analyses (Legewie, 2013). Here, we provide additional evidence on how the expression of hate is connected to the uncertainty about social norms. Since anti-hate norms play an important role in containing the expression of prejudice, understanding how terrorist attacks may impact the strength of the social norm is essential to understanding many responses to terrorist attacks.

Our empirical study compares the levels of pre- and post-attacks hate speech in contexts with different strengths of the descriptive anti-hate norm. This allows us to investigate context effects on the comments posted in our experimental online forum. The design also allows us to disentangle these effects from the impact of the attacks on individual attitudes. Additionally, we investigate the effect of the terrorist attacks on normative uncertainty by comparing the effect of descriptive norms after the attacks across all conditions and topics.

We find that the hostility of hate speech increases after Islamist terrorist attacks only towards refugees, and only if previous anti-refugee comments signal the reduced validity of anti-hate norms. In all other conditions, that is, on other topics or when negative comments are absent, hate speech either does not change or is even slightly reduced when only positive comments are shown, even though this reduction is far from statistical significant. The effect of the terrorist attacks on hate speech is thus highly dependent on the local context and the respective norms therein. It also suggests that in natural environments with no external intervention, Islamist terrorist attacks might indeed increase the level of online hate speech against refugees. It is worth noting that our estimation of the absolute effect might be an underestimation of the true effect. We used the same comments, and therefore the same level of hate speech before and after the attack in our mixed condition. In a natural environment, one would expect that comments would be more negative after the attack, and consequently the reaction to these comments should be even more negative than found in our experiment. This could create a reinforcing loop and thus lead to an escalation of hostility.

Online hate against refugees increases after the attacks both compared with levels of hate against refugees before the attacks, and also relative to the increase towards gender rights. This increase is not found when the anti-hate norm is exogenously manipulated to remain strong. Under the fairly innocent assumption of proper randomization of attitudes into experimental conditions, this difference can be solely attributed to the interplay between normative context and the normative uncertainty created by the attacks. Our results show that norm conformity is larger after the terrorist attacks in topics linked to the events. Our results thus imply that attitudinal changes due to terrorist attacks are more likely to be voiced if the perceived social acceptability of expressing prejudice increases, and that normative uncertainty is an endogenous consequence of terrorist attacks.

The terrorist attacks analyzed here were carried out in a very particular historical context. They happened at the peak of a domestic political crisis due to a sudden influx of refugees, in which immigration laws, integration, and the threat of Islamist terrorism were widely discussed in German media. Does this limit the scope conditions of our theory? While it is impossible to give a definite answer in the absence of additional empirical data, we believe that the proposed mechanism's scope may be much more general than that. It could, for instance, apply to other instances of external shocks, e.g. the surprising election of political candidates, which could alter our beliefs about the expectations of others and rapidly change a previously shared consensus. Bursztyn et al. (2017) show in a donation experiment that monetary transfers to people stating anti-immigrant opinions increased after the election of Donald Trump in 2016. Along these lines, Crandall et al. (2018) suggest that the election also weakened the norm of expressing prejudice for the specific groups targeted by the election campaign. The same could hold for the erosion of anti-hate norms in the public space more generally, particularly if prominent actors publicly break the norms. Our mechanism of increased receptiveness to social cues could also apply to other challenged norms. We would, for instance, predict that an unexpected court ruling concerning gay marriage or public scandals of underage drinking could create similar patterns of local anomie. Finally, descriptive norms might not be the only cue that could create this effect. Álvarez-Benjumea and Winter (2018) provide suggestive evidence that prescriptive norms could be signaled via instances of sanctioning hate speech. These sanctions could then serve as a cue for normative behavior.

Furthermore, the effect of changing social norms might also vary over time. We have described a mechanism whereby social norms regulate the expression of attitudes; in the long run, however, norms might actually change personal attitudes (Stangor et al., 2001), or our perception of how prejudiced we are in comparison with others (Crandall et al., 2018). Our results support previous empirical findings that show how uncertainty more generally makes people receptive to social influence processes, such as herding behavior (Bikhchandani et al., 1992), conformist influence

(Centola, Willer, & Macy, 2005), and norm conformity (Deutsch & Gerard, 1955). We thus clarify specific domains where this is the case.

Our findings have direct implications for sociological research methods. Legewie (2013) stresses the importance of dramatic events in one country as a potential source of severe bias in cross-national research. Differences in attitudes between states could be largely driven by the temporal proximity of short-lived attitude changing events. Our findings imply that the particular social context in which the data are collected should also be taken into account because the perceived normativity of certain behaviors in the local context represents an additional source of bias.

It is worth noting that the duration of the effect of the terrorist attacks on normative uncertainty is unclear, and that this effect could change over time. That said, our results suggest that supervisors of public discussions, either in the virtual domain or the real world, may be well advised to implement measures to ensure a well tempered atmosphere. A few bad examples could already lead to an erosion of decency norms. As an unintended consequence, this may lead to the (self-)exclusion of marginalized groups from the discussion, and in the worst case to a breakdown of a whole debate. Naturally, this moderation can be a delicate matter, and the acceptance of these measures may depend to a large extent on existing cultural expectations. It should therefore, if at all, be applied with a strong sense of proportion.

As a last remark, we would like to point out that we obviously did not initially plan to study the effects of terrorist attacks, but were confronted with it by a tragic coincidence. We acknowledge that a better design to test every path of our theory, and to discriminate between our argument and competing explanations, would have looked much more comprehensive. We could have measured attitudes towards immigrants and gender rights, sociodemographic characteristics of the respondents, media consumption before, between and after the attacks, and maybe anomie. Even more convincing would have been a design in which we would have been able to collect twice the amount of data from the same participants before and after the attacks to estimate individual reactions. But all this would have required us to know about these events in advance, which we obviously did not.

Appendices

Appendix A

Appendix Chapter 1

A.1 Appendix A: Materials

A.1.1 Pre-experimental Work

Keywords Used in Pictures Search

The images were obtained from Twitter and Google images in March 2016. We used different tags or keywords to search for images that were twitted using them. We used both German and English terms that are often used on German social media.

Sharia, Multiculturalism, Terrorism, Religion, Transgender, Gay, Homosexuality, Sexism, Discrimination, Refugees, *aufschrei*, Sexism, Immigration, homosexuality, *Einwanderung*, Diversity, Queer, Begging, Atheism, islamization, Religion, #tolerist

Online Survey

The following questionnaire was answered by 90 participants during the pre-experimental phase.

Survey (*In German in the original*)

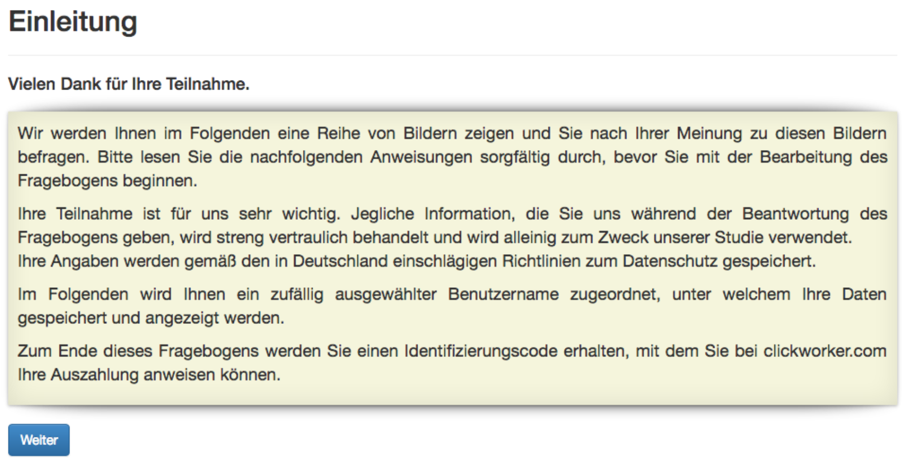
1. How controversial do you think the picture above is? (on a scale from 1 to...)

2. Would you say this picture represent more a positive or negative view of the issue depicted in the image?

3. Which topic would you say the picture above represents best? (Choose only one)

Refugees, Immigration, *Gendermainstreaming*, Transgender, Genderbender, #aufschrei, Aggressive behavior, Romany, Begging, Grexit, Eurocrisis, Zionism or Judaism

A.1.2 Experimental Instructions



Einleitung

Vielen Dank für Ihre Teilnahme.

Wir werden Ihnen im Folgenden eine Reihe von Bildern zeigen und Sie nach Ihrer Meinung zu diesen Bildern befragen. Bitte lesen Sie die nachfolgenden Anweisungen sorgfältig durch, bevor Sie mit der Bearbeitung des Fragebogens beginnen.

Ihre Teilnahme ist für uns sehr wichtig. Jegliche Information, die Sie uns während der Beantwortung des Fragebogens geben, wird streng vertraulich behandelt und wird alleinig zum Zweck unserer Studie verwendet. Ihre Angaben werden gemäß den in Deutschland einschlägigen Richtlinien zum Datenschutz gespeichert.

Im Folgenden wird Ihnen ein zufällig ausgewählter Benutzername zugeordnet, unter welchem Ihre Daten gespeichert und angezeigt werden.

Zum Ende dieses Fragebogens werden Sie einen Identifizierungscode erhalten, mit dem Sie bei clickworker.com Ihre Auszahlung anweisen können.

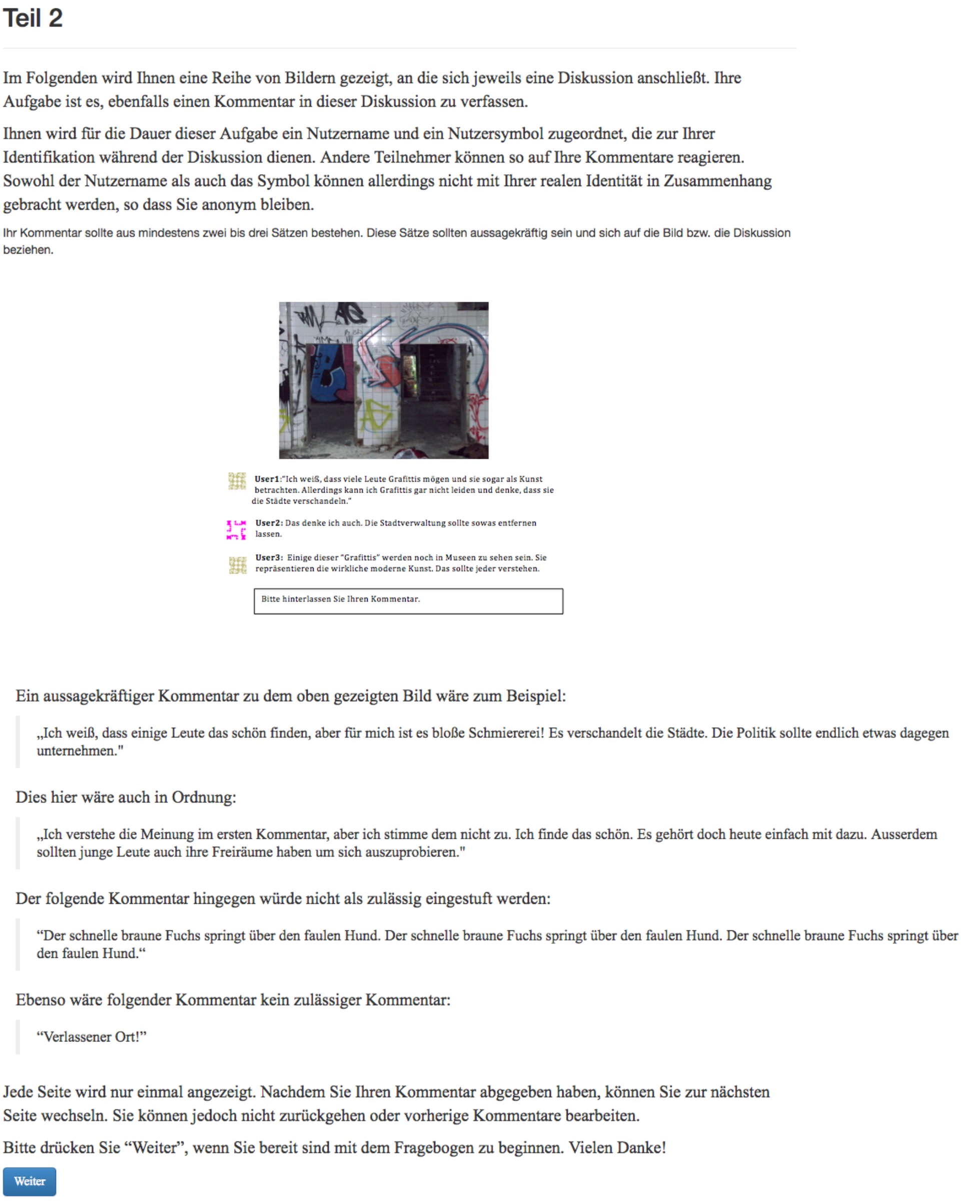
Figure A1: Introduction Page of the Experiment.

Introduction (In English)

Thank you for your participation.

We will show you a series of pictures and ask you comment on them. Please read the following instructions carefully before you begin the task. Your participation is very important to us. Any information you provide to us during the task will be strictly confidential and will be used solely for the purpose of our study. Your data will be stored in accordance with the relevant data protection guidelines in Germany.

You will be assigned a random user name, and your input will be stored and displayed under this username. At the end, you will be given an identification code, which will allow you to claim your payment at `clickworker.com`.



Teil 2

Im Folgenden wird Ihnen eine Reihe von Bildern gezeigt, an die sich jeweils eine Diskussion anschließt. Ihre Aufgabe ist es, ebenfalls einen Kommentar in dieser Diskussion zu verfassen.

Ihnen wird für die Dauer dieser Aufgabe ein Nutzername und ein Nutzersymbol zugeordnet, die zur Ihrer Identifikation während der Diskussion dienen. Andere Teilnehmer können so auf Ihre Kommentare reagieren. Sowohl der Nutzername als auch das Symbol können allerdings nicht mit Ihrer realen Identität in Zusammenhang gebracht werden, so dass Sie anonym bleiben.

Ihr Kommentar sollte aus mindestens zwei bis drei Sätzen bestehen. Diese Sätze sollten aussagekräftig sein und sich auf die Bild bzw. die Diskussion beziehen.

User1:"Ich weiß, dass viele Leute Grafittis mögen und sie sogar als Kunst betrachten. Allerdings kann ich Grafittis gar nicht leiden und denke, dass sie die Städte verschandeln."

User2: Das denke ich auch. Die Stadtverwaltung sollte sowas entfernen lassen.

User3: Einige dieser "Grafittis" werden noch in Museen zu sehen sein. Sie repräsentieren die wirkliche moderne Kunst. Das sollte jeder verstehen.

Bitte hinterlassen Sie Ihren Kommentar.

Ein aussagekräftiger Kommentar zu dem oben gezeigten Bild wäre zum Beispiel:

> „Ich weiß, dass einige Leute das schön finden, aber für mich ist es bloße Schmiererei! Es verschandelt die Städte. Die Politik sollte endlich etwas dagegen unternehmen."

Dies hier wäre auch in Ordnung:

> „Ich verstehe die Meinung im ersten Kommentar, aber ich stimme dem nicht zu. Ich finde das schön. Es gehört doch heute einfach mit dazu. Ausserdem sollten junge Leute auch ihre Freiräume haben um sich auszuprobieren."

Der folgende Kommentar hingegen würde nicht als zulässig eingestuft werden:

> "Der schnelle braune Fuchs springt über den faulen Hund. Der schnelle braune Fuchs springt über den faulen Hund. Der schnelle braune Fuchs springt über den faulen Hund.“

Ebenso wäre folgender Kommentar kein zulässiger Kommentar:

> "Verlassener Ort!"

Jede Seite wird nur einmal angezeigt. Nachdem Sie Ihren Kommentar abgegeben haben, können Sie zur nächsten Seite wechseln. Sie können jedoch nicht zurückgehen oder vorherige Kommentare bearbeiten.

Bitte drücken Sie "Weiter", wenn Sie bereit sind mit dem Fragebogen zu beginnen. Vielen Danke!

Figure A2: Instructions of the Experiment.

Instructions (In English)

You will see a series of pictures with a discussion below. Your task is to join the discussion on the topic(s) depicted in the picture(s). Please write at least two to three sentences per discussion. These sentences should be meaningful and relate to the picture/discussion.

A valid comment on the discussion above would be:

"I know that some people like them and even consider them to be art. However, I really dislike graffiti or "street art" and some call it. I think it impoverishes the way a city looks"

"I do understand the opinion in comment 1, although I pretty much disagree. Most of the places that are now covered by graffiti were previously abandoned and looked very dirty and ugly already"

The following comment would not count as valid:

"The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog."

The following is also not a sufficient comment:

"Abandoned place"

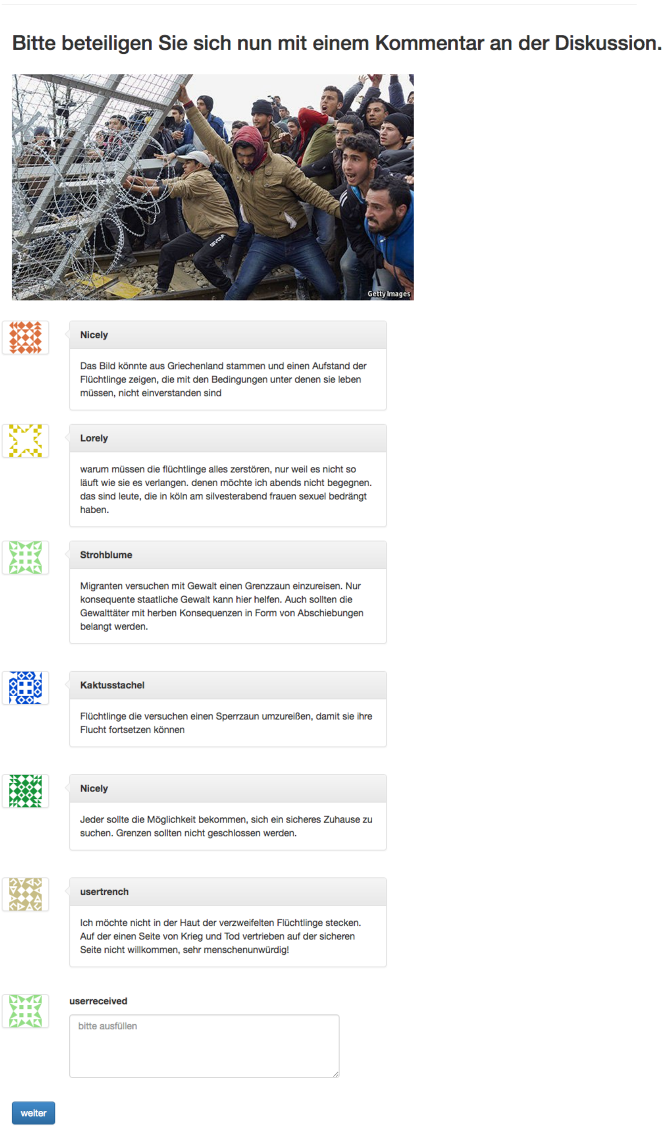
Each page will be shown just once. Once you have finished with your comment you can go to the next page, but you cannot go back or edit previous comments.

Please press the "start" button once you are ready to start the survey.

Thanks!

A.1.3 Screenshots of the Online Experimental Forum

In this section, screenshots of the different experimental treatments as they are shown in the online forum during the experiment are depicted.



Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.

Nicely

Das Bild könnte aus Griechenland stammen und einen Aufstand der Flüchtlinge zeigen, die mit den Bedingungen unter denen sie leben müssen, nicht einverstanden sind

Lorely

warum müssen die flüchtlinge alles zerstören, nur weil es nicht so läuft wie sie es verlangen. denen möchte ich abends nicht begegnen. das sind leute, die in köln am silvesterabend frauen sexuel bedrängt haben.

Strohblume

Migranten versuchen mit Gewalt einen Grenzzaun einzureisen. Nur konsequente staatliche Gewalt kann hier helfen. Auch sollten die Gewalttäter mit herben Konsequenzen in Form von Abschiebungen belangt werden.

Kaktusstachel

Flüchtlinge die versuchen einen Sperrzaun umzureißen, damit sie ihre Flucht fortsetzen können

Nicely

Jeder sollte die Möglichkeit bekommen, sich ein sicheres Zuhause zu suchen. Grenzen sollten nicht geschlossen werden.

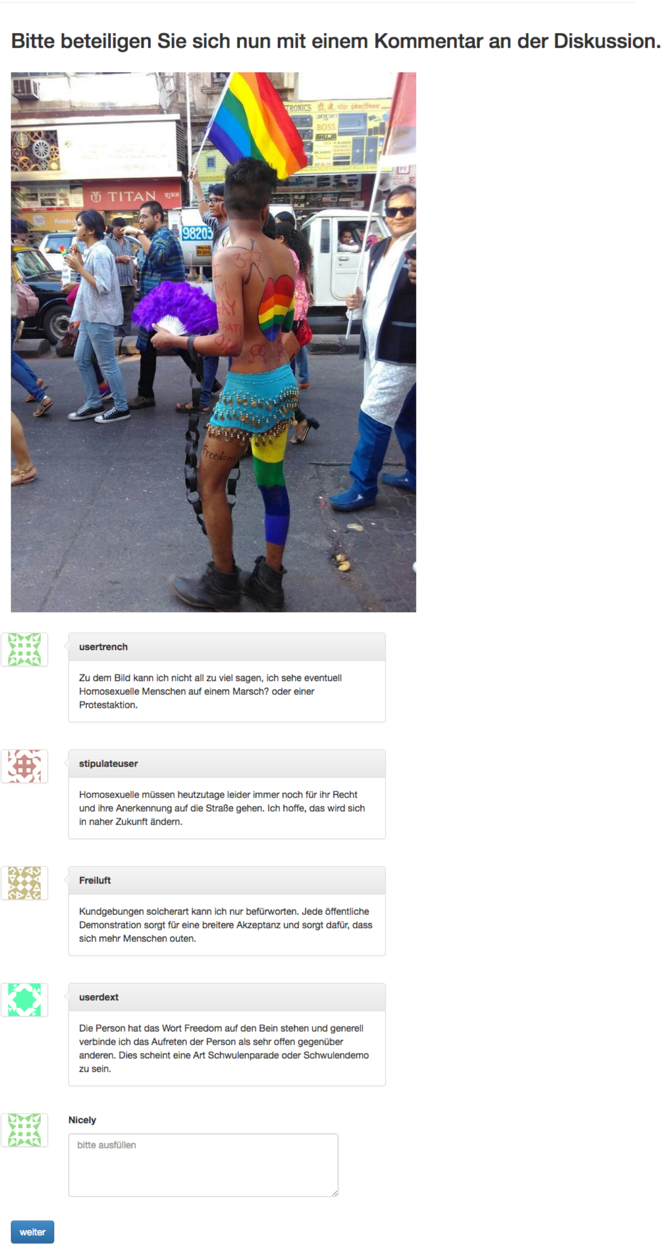
usertrench

Ich möchte nicht in der Haut der verzweifelten Flüchtlinge stecken. Auf der einen Seite von Krieg und Tod vertrieben auf der sicheren Seite nicht willkommen, sehr menschenunwürdig!

userreceived

bitte ausfüllen

Figure A3: Screenshot of the Baseline treatment as shown in the online forum.



Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.

usertrench

Zu dem Bild kann ich nicht all zu viel sagen, ich sehe eventuell Homosexuelle Menschen auf einem Marsch? oder einer Protestaktion.

stipulateuser

Homosexuelle müssen heutzutage leider immer noch für ihr Recht und ihre Anerkennung auf die Straße gehen. Ich hoffe, das wird sich in naher Zukunft ändern.

Freiluft

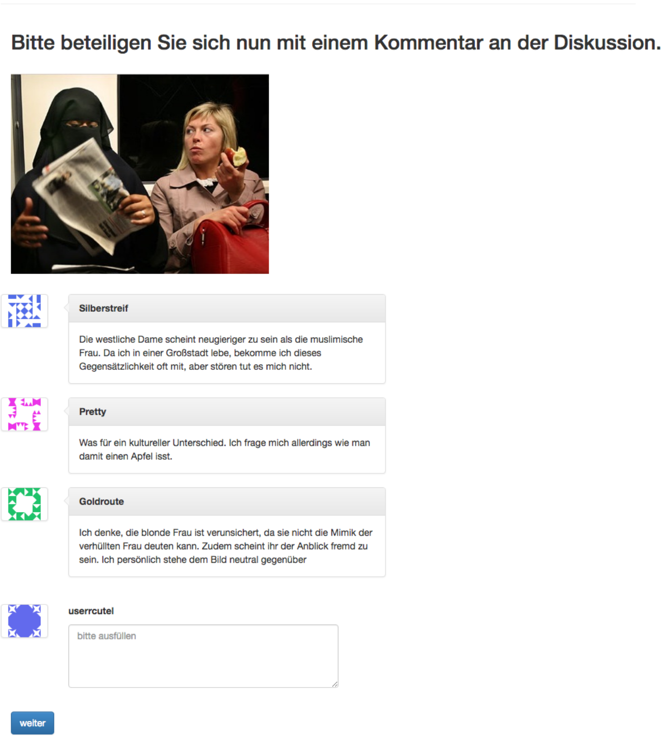
Kundgebungen solcherart kann ich nur befürworten. Jede öffentliche Demonstration sorgt für eine breitere Akzeptanz und sorgt dafür, dass sich mehr Menschen outen.

userdext

Die Person hat das Wort Freedom auf den Bein stehen und generell verbinde ich das Aufreten der Person als sehr offen gegenüber anderen. Dies scheint eine Art Schwulenparade oder Schwulendemo zu sein.

Nicely

Figure A4: Screenshot of the Censored treatment as shown in the online forum.



Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.

Silberstreif

Die westliche Dame scheint neugieriger zu sein als die muslimische Frau. Da ich in einer Großstadt lebe, bekomme ich dieses Gegensätzlichkeit oft mit, aber stören tut es mich nicht.

Pretty

Was für ein kultureller Unterschied. Ich frage mich allerdings wie man damit einen Apfel isst.

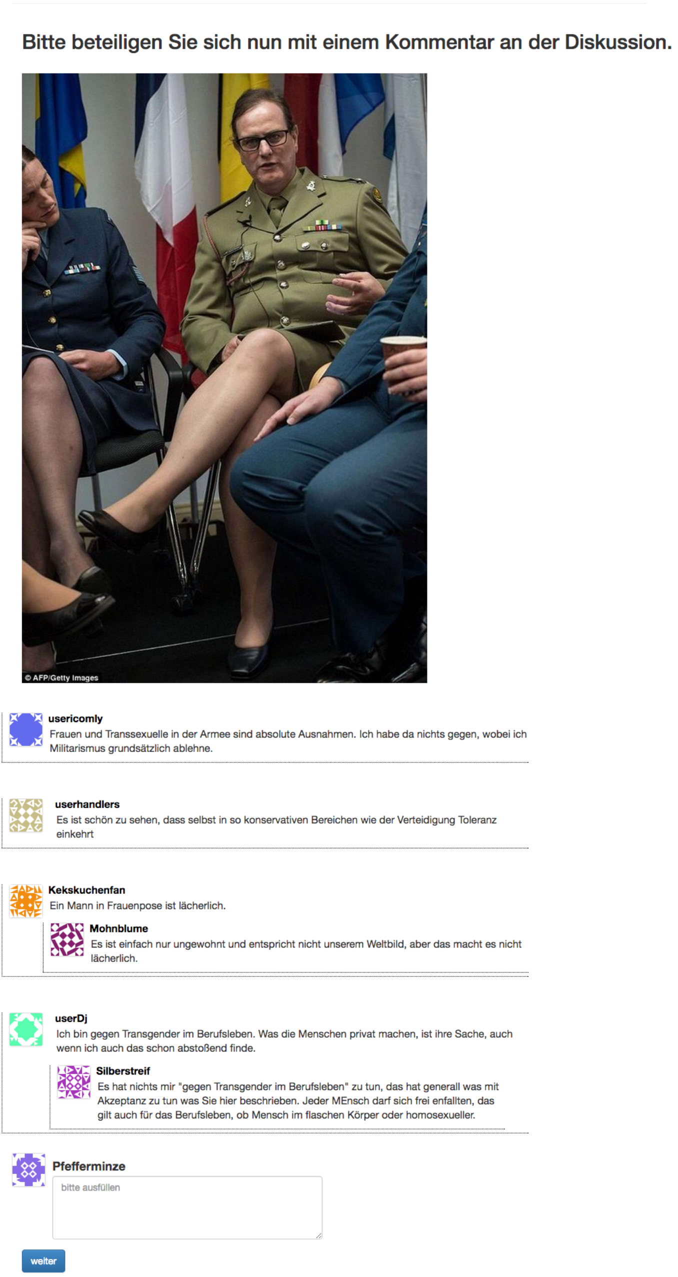
Goldroute

Ich denke, die blonde Frau ist verunsichert, da sie nicht die Mimik der verhüllten Frau deuten kann. Zudem scheint ihr der Anblick fremd zu sein. Ich persönlich stehe dem Bild neutral gegenüber

userrcutel

bitte ausfüllen

Figure A5: Screenshot of the Extremely Censored treatment as shown in the online forum.



Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.

usericomly
Frauen und Transsexuelle in der Armee sind absolute Ausnahmen. Ich habe da nichts gegen, wobei ich Militarismus grundsätzlich ablehne.

userhandlers
Es ist schön zu sehen, dass selbst in so konservativen Bereichen wie der Verteidigung Toleranz einkehrt

Kekskuchenfan
Ein Mann in Frauenpose ist lächerlich.

Mohnblume
Es ist einfach nur ungewohnt und entspricht nicht unserem Weltbild, aber das macht es nicht lächerlich.

userDj
Ich bin gegen Transgender im Berufsleben. Was die Menschen privat machen, ist ihre Sache, auch wenn ich auch das schon abstoßend finde.

Silberstreif
Es hat nichts mir "gegen Transgender im Berufsleben" zu tun, das hat generall was mit Akzeptanz zu tun was Sie hier beschrieben. Jeder MEnsch darf sich frei enfallten, das gilt auch für das Berufsleben, ob Mensch im flaschen Körper oder homosexueller.

Pfefferminze
bitte ausfüllen

Figure A6: Counter-speaking treatment as shown in the online forum.

A.1.4 Examples of Comments

Hate comments

- Comment 1261 (User 171): "Flüchtlingskrise. Die können weiter wandern weg von europa. das sind nicht nur kriegsflüchtlinge, zu 90 prozent sind es nur sozialschmarotzer die hier machen können was die wollen." *(Refugee crisis. They can continue walking away from Europe. They are not just war refugees, 90 per cent are nothing but social parasites who can do whatever they want here.)*(Hate score: 8.66)
- Comment 1029 (User 112): "Schwule Kerle sind das Letzte, was ich dulden würde. Schon gar nicht in der Öffentlichkeit." *(Gay guys are the last thing I would tolerate, especially not in public.)*(Hate score:9)

Neutral comments

- Comment 892 (User 179): "Generell bin ich dagegen sich in der Öffentlichkeit wild zu küssen. Aber gegen einen Kuss habe ich nichts."*(In general I am against passionate kissing in public. But I have nothing against a kiss.)* (Hate score: 5)

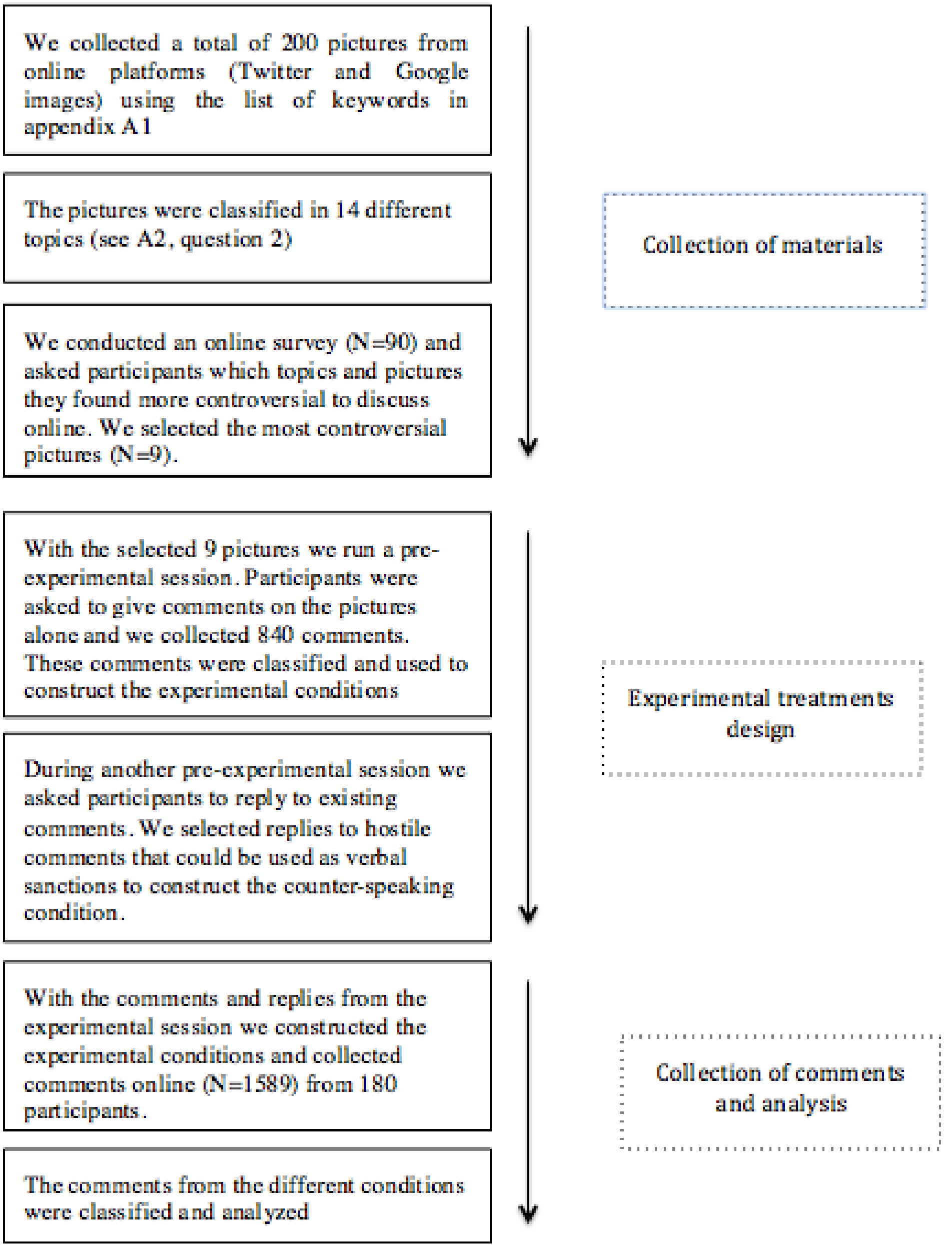
Friendly comments

- Comment 1572 (User 71): "Das ist wirklich eine wunderschöne Szene. Es sieht nach einer absolut glücken Familie aus. Wahrscheinlich sind sie glücklicher als so manches hetero Paar."*(This is really a wonderful scene. It looks like an absolutely happy family. They are probably happier than many heterosexual couples.)*(Score of 1,66)
- Comment 216 (User 63): "Super Daumen hoch für diese Leute die den Mumm haben sich der Ignoranz zu stellen." *(Super thumbs up for these people, who have the guts to face up to this ignorance.)*(Score of 2)

A.1.5 Timeline of the Experiment

The process is divided into three phases: collection of materials, treatment design, and collection of comments and analysis. All pre-experimental sessions were conducted using workers recruited from Clickworker. During the collection of materials (March to April, 2016), we collected images using the keywords in A1. Pictures were collected from Google images and Twitter. The pictures were classified into different categories (see categories in question 3 of the online survey in A2). In April 2016, we ran an online survey (N=90) with a sample of German speakers from Clickworker and selected the pictures and topics rated as "more controversial". In June 2016 we made the forum with the 9 selected pictures available online to workers. The comments collected in this first pre-experimental session were rated by three independent raters using the hate score (1 to 9 scale), and classified into 3 categories: friendly, neutral, and hostile. In a second pre-experimental session using workers, we collected replies to hostile comments, which could be used as verbal sanctions to construct the counter-speaking condition. We used them to construct the counter-comments condition. All experimental conditions are constructed using previous comments. The exact comments and the order of appearance in the discussion were automatically selected each time a participant joined the experiment, e.g. when a participant was allocated to the extremely censored condition she could see 3 randomly chosen friendly comments.

Finally, we conducted the experiment during the 4-5th August 2016, we made our forum available online and distributed the link to the participants in the experiment (N=180) via Clickworker. The raters then classified the collected comments during a period of 2 weeks. The score and items were finally analyzed.



We collected a total of 200 pictures from online platforms (Twitter and Google images) using the list of keywords in appendix A1

The pictures were classified in 14 different topics (see A2, question 2)

We conducted an online survey (N=90) and asked participants which topics and pictures they found more controversial to discuss online. We selected the most controversial pictures (N=9).

With the selected 9 pictures we run a pre-experimental session. Participants were asked to give comments on the pictures alone and we collected 840 comments. These comments were classified and used to construct the experimental conditions

During another pre-experimental session we asked participants to reply to existing comments. We selected replies to hostile comments that could be used as verbal sanctions to construct the counter-speaking condition.

With the comments and replies from the experimental session we constructed the experimental conditions and collected comments online (N=1589) from 180 participants.

The comments from the different conditions were classified and analyzed

Collection of materials
Experimental treatments design
Collection of comments and analysis

Figure A7: Timeline of the Experiment.

A.2 Appendix B: Further Analyses

A.2.1 Analyses of Extremely Hateful Comments

We estimated linear conditional quantile regressions at the 0.75, 0.90, 0.95 and 0.99 quantiles of the hate score distribution. The table shows the treatment effects estimates for each of the quantiles or quantile coefficients. This gives a more complete description of how the effect of the treatments work at the higher quantiles, which correspond to the most hostile comments, of the conditional distribution of the hate score Table A1.

In Table A2 we present three further models of hostile comments (more than 7 in the hate score). Model 1 shows the results for a logistic regression model with clustered standard errors at the individual level (180 clusters). Model 2 shows a logistic regression model with bootstrapped clustered standard errors. The replications in model 2 are based on 180 clusters in the data. Both models show that the effect of the censored treatment is robust, even after taking into account the nested structure of the comments. Model 3 is a rare events logistic regression model. This model, which corrects for small-sample bias, also supports the reduction in the number of hostile comments in the censored treatment.

Table A1: Results from the linear conditional quantile regression.

	Coef. (Std. Error)
Quantile 0.75	
Constant	5.33 (0.13)***
Counter-speaking	−0.33 (0.14)**
Censored	−0.33 (0.14)**
Extremely censored	−0.33 (0.18)†
Quantile 0.90	
Constant	6.33 (0.09)***
Counter-speaking	0.00 (0.12)
Censored	−0.67 (0.20)
Extremely censored	−0.33 (0.20)
Quantile 0.95	
Constant	7.00 (0.20)***
Counter-speaking	−0.33 (0.36)
Censored	−0.67 (0.23)
Extremely censored	−0.33 (0.23)
Quantile 0.99	
Constant	8.33 (0.42)***
Counter-speaking	0.00 (0.59)
Censored	−1.33 (0.50)**
Extremely censored	−1.00 (0.44)*

Notes:: Linear conditional quantile regressions at the 0.75, 0.90, 0.95 and 0.99 quantiles of the hate score distribution. Quantile regression coefficients are listed with standard errors in parentheses (Obs=1469). Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test

The estimate for the extremely censored condition is only significantly larger than the estimate for the censoring treatment at p < 0.10.

Table A2: Logistic regression results with clustered standard errors (model 1), logistic regression with bootstrapped clustered standard errors (model 2), and Rare Events regression (model 3) of hostile comments (more than a 7 in the hate score).

	Model (1)	Model (2)	Model (3)
Constant	-2.58 (0.36)***	-2.68 (0.41)***	-2.66 (0.21)***
Counter-speaking	-0.30 (0.50)	-0.30 (0.60)	-0.29 (0.32)
Censored	-1.85 (0.70)**	-1.85 (0.73)*	-1.75 (0.64)**
Extremely censored	-0.30 (0.50)*	-0.41 (0.73)	-0.40 (0.34)
Log Likelihood	-245.13	-245.13	
AIC	498.26	498.26	498.26
Obs.	1469	1469	1469
Pseudo-R^2	0.0341	0.0341	

Notes: Notes: Logistic regression with clustered s.e at the individual level (Model 1), logistic regression with bootstrapped clustered s.e at the individual level (Model 2), and rare events logistic regression (Model 3) with hostile comments as dependent variable. Standard errors in parentheses. Replications of model 2 based on 180 clusters in the data. One or more parameters could not be estimated in 24 bootstrap replicates; standard-error estimates in Model 2 include only complete replications. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test

A.2.2 Robustness Checks

Table A3 shows the results of rare events logistic regression of the comments classified as a norm violation. The results from the multilevel random model in Table 6 are robust. The results from this model have to be interpreted with caution because the rare events logistic model does not account for the nested structure of the data.

Table A3: Rare events analysis of norm-violations.

	Model (1)
Constant	-1.91 (0.16)***
Counter-speaking	0.08 (0.22)
Censored	-0.95 (0.28)**
Extremely censored	-0.50 (0.25)*
AIC	943.07
Obs.	1469

Notes: Rare events logistic regression estimates (Model 1) and logistic regression with clustered standard errors (Model 2) with norm violations as the dependent variable. Standard errors in parentheses. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test

Table A4 shows the results from a multilevel model with two random intercepts (picture and subject) similar to models in table 4. The model shows the interaction effects of the treatments and topic combined after the simple effects for topics and treatments have been taken into account. Poverty is used as the reference category, which means that the interaction terms have to be understood as compared with that category. This model is a reparametrization of model 3 in Table 4.

We also validated the model of hate score using a jackknife “leave-one-out’” resampling technique. We computed the main model (model 1 in Table 1.4) leaving out one subject at a time, with a total of 180 models. The distribution of the coefficient estimates of the treatments using this strategy are presented here. Figure Figure A8 shows the estimates for the counter-speaking treatment, the censored treatment,

	Model 1
Main effects	
Constant	4.20 (0.37)**
Counter-speaking	0.07 (0.24)
Censored	−0.02 (0.24)
Extremely censored	−0.12 (0.25)
LGTB	0.22 (0.44)
Refugees/Multiculturality	0.97 (0.41)*
Feminism	0.19 (0.41)
Interaction effects	
Poverty×Counter-speaking	
LGTB×Counter-speaking	−0.24 (0.25)
Refugees×Counter-speaking	−0.33 (0.24)
Feminism×Counter-speaking	−0.18 (0.23)
Poverty×Censored	
LGTB×Censored	−0.31 (0.25)
Refugees×Censored	−0.62 (0.24)**
Feminism×Censored	−0.30 (0.23)
Poverty×Extremely	
LGTB×Extremely	−0.33 (0.26)
Refugees×Extremely	−0.48 (0.25)$^{\cdot}$
Feminism×Extremely	−0.16 (0.24)
AIC	4371.48
BIC	4472.04
Log Likelihood	-2166.74
Obs	1469
Groups:Subjects	180
Var:Subjects	9
Groups:Subjects	0.44
Var:Subjects	0.11
Residual Variance	0.90

Notes: Linear mixed model fit by REML. Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for models of hate speech score. The table lists mean regression coefficient estimates with standard errors in parentheses and p-values calculated based on Satterthwate's approximations. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test

Table A4: Results from a multilevel model with two random intercepts (picture and subject).

and the estimates for the extremely censored treatment. The value of the coefficients does not change significantly after with the removal

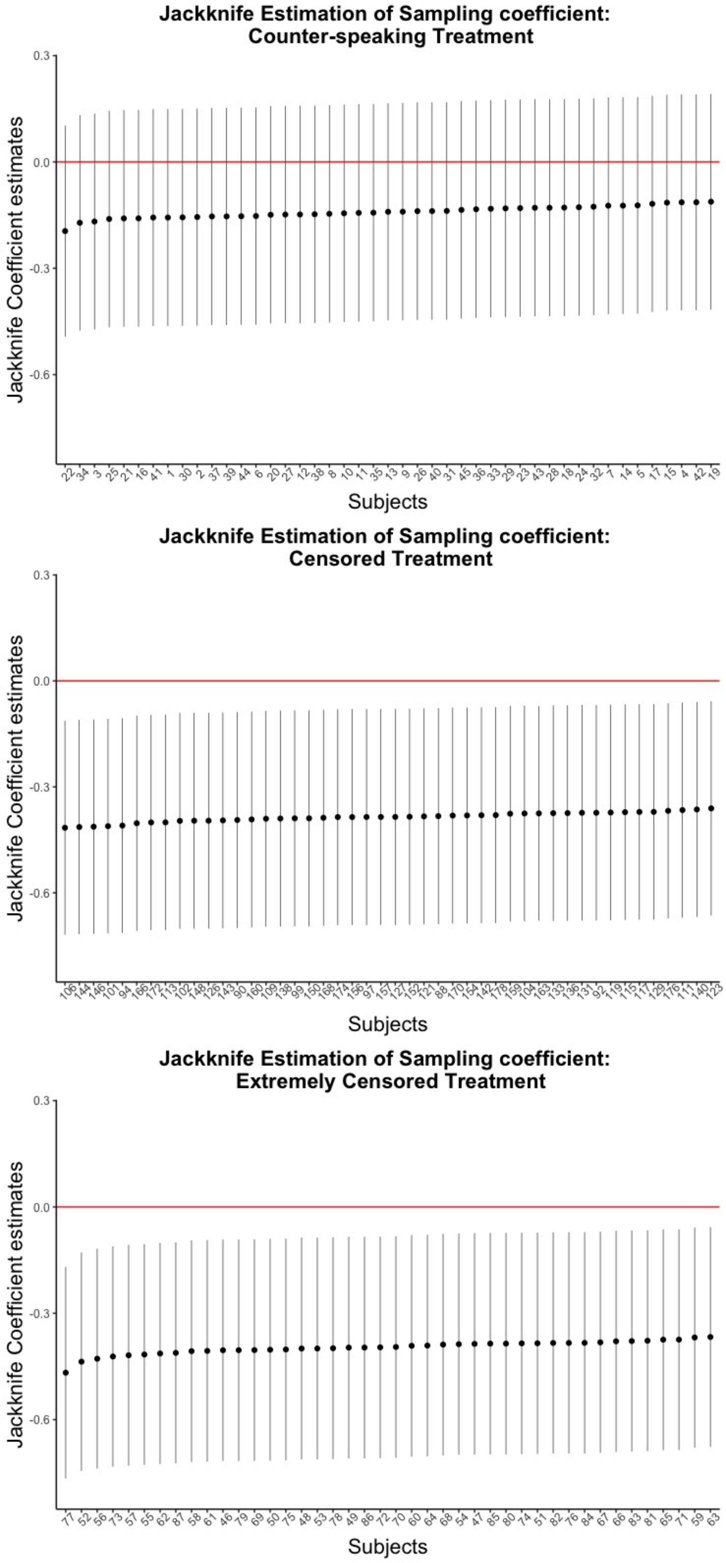


Figure A8: Jackknife Estimation of Sampling Coefficients.

Appendix B

Appendix Chapter 2

B.1 Appendix A: Materials

B.1.1 Sampling Strategy

Participants were recruited from a crowdsourcing Internet marketplace.[1] The experiment was conducted entirely in German and the sample was restricted to residents in Germany. The demographic characteristics of the participants are depicted in Table 2.3. Participants were told that they were taking part in an experimental study, but not the actual purpose of the experiment. Participants were explained the basic functioning of the forum and asked to join the conversations and leave comments on the different posts. They were also told that they were going to be asked about a donation decision to an organization that would be randomly selected for them. In order to ensure external validity of the experiment various measures were taken: i) participants remained anonymous and no personal data were recorded, ii) participation was voluntary, and iii) payment did not depend on performance.

At the beginning of the experiment, they were given a user name and an avatar. Once the experiment started, every participant was consecutively presented with the discussions and asked to leave a comment at the bottom of each thread. Each participant was required to leave a comment in each forum page, with a total of five comments per participant. The donation decision page popped up only after one of the forum pages. The time of the decision was randomly selected for each participant at the beginning of the experiment with probability for each of the possible 5 time. At the end of the experiment, participants completed the demographic questions and received a code that could be exchanged for the payment in an anonymously.

Selection of the Organizations

Respondents of the pre-experimental survey were given a list with nine pre-selected organizations. They were asked whether they knew each specific organization (*From the list of associations below, please mark those that you have heard of*) and to classify them as to whether they support refugees and immigration or, on the contrary, have an anti-immigrant agenda (*Please classify the following organizations according to whether they are supportive or hostile to immigrants. Give a score between 1 and 100, where 1 means refugee-friendly and 100 anti-refugee*).

[1] www.clickworker.com

Table A1: Organizations listed in the pre-experimental survey. Percentage of respondents who had previously heard of each organization, and the average anti-refugee score (1-100).

Organization	Known by (%)	Anti-refugee Score (1-100)
AfD	31%	81
Identitare bewegung	11%	65
Buergerbewegung PRO-NRW	12%	66
Mensch Mensch Mensch	1%	25
Fluechtlinge Wilkommen	6%	19
Pro Asyl	21%	17
Gustav Stresemann Stiftung	7%	47
Desiderius Erasmus Stiftung	3%	46
Amadeu Antonio Stiftung	6%	35

B.1.2 Construction of the Forum

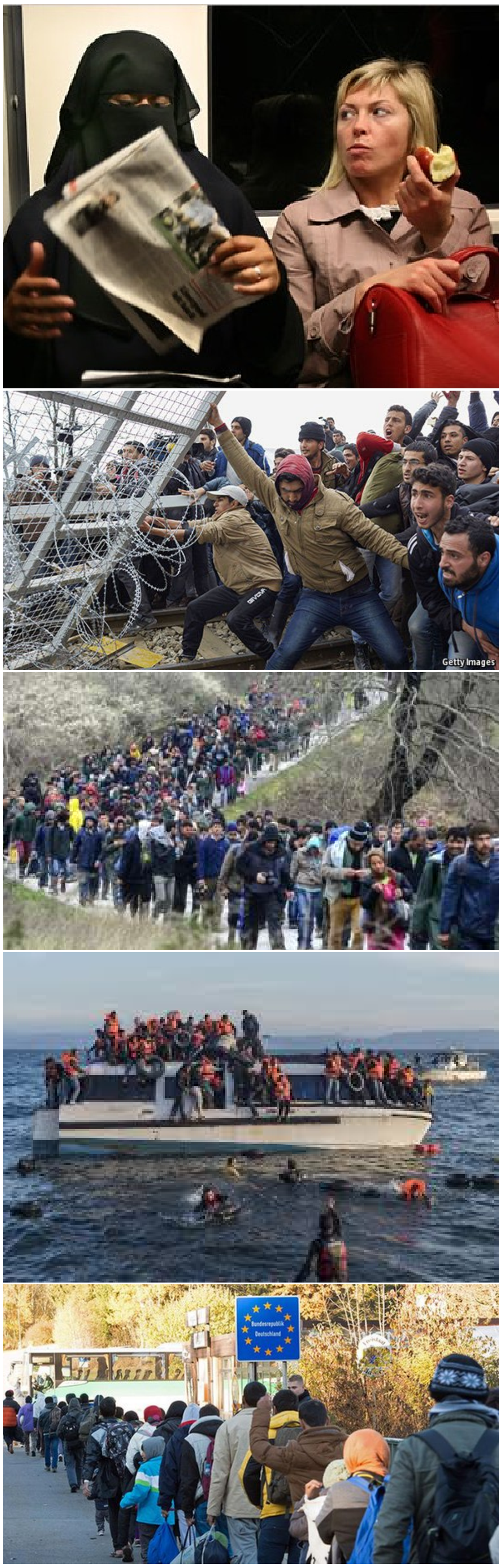


Figure A1: Pictures used in the construction of the experiment (1 to 5).

B.1.3 Experimental Instructions

Introduction

Please read the following instructions carefully before continuing to the survey. These instructions will be shown only once. Your participation is very important for us. All the information you give during our survey will be treated in strict confidentiality and used only for the purpose of this study. You will be assigned a random user name, and your input will be stored and displayed under this username. After receiving your payment, your Clickworker account cannot be tied to the information you gave in the survey.

Instructions for the Participation in the Forum

Welcome to this online forum. You will see a series of pictures with a discussion below. Your task is to join the discussion on the topic(s) depicted in the picture(s).

Please write at least one to two sentences per discussion. These sentences should be meaningful and relate to the picture/discussion.

A valid comment on the discussion above would be:

"I know that some people like them and even consider them to be art. However, I really dislike graffiti or "street art" and some call it. I think it impoverishes the way a city looks"

"I do understand the opinion in comment 1, although I pretty much disagree. Most of the places that are now covered by graffiti were previously abandoned and looked very dirty and ugly already"

The following comment would not count as valid:

"The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog"

The following is also not a sufficient comment:

"Abandoned place"

Each page will be shown only once. Once you have finished with your comment you can go to the next page, but you cannot go back or edit previous comments.

Additionally, we will ask you to make a decision about a donation to randomly selected German non-profit associations. The decision page will pop up while you are discussing in the forum. After the decision, you may continue discussing.

Please press the "start" button once you are ready to start.

Thanks!

Debriefing Message

END!!!!

Thank you very much for your participation. Your code for claiming the payment in Clickworker is: xxxxxx

x% of the participants in the forum until now have decided to donate (your decision is not part of this count).

This study intends to test the effect that online participation has on your likelihood of revealing your political opinions. Thank you very much again for your participation.

List of deleted comments

9QJD0IME
gg
rf
sdvgsfdgsdf
e<fsfsefse
http://134.76.24.117/static/donation_experiment/images/avatars/20.91e9c6d4e8eb.png
_
test
clickworker test
Traurig
dfgrdgrsgd
dasdas
adsadasd
Test
9QJD0IME
gg
uu9
dsfsdff
esfsefse
http://134.76.24.117/static/donation_experiment/images/avatars/10.450aeba94eeb.png
_
test
clickworker test
a
werwerewrwer
sadasd
asdasdasd
Test
9QJD0IME
gg
u9nu
sdfsdfsdf
sefsefesf
http://134.76.24.117/static/donation_experiment/images/avatars/10.450aeba94eeb.png
.
test
clickworker test
d & ergreerger
asdas
asdsada
Test
9QJD0IME
gg
j9uniu
sdfdsfsdf
esfseffe
http://134.76.24.117/static/donation_experiment/images/avatars/10.450aeba94eeb.png
.
test
clickworker test

d
ergerregeg
dasd
adasdassad
test
9QJD0IME
gg
uninu
sdfdsfsdf
sefsefsef
http://134.76.24.117/static/donation_experiment/images/avatars/10.450aeba94eeb.png
.
test
clickworker test
d
ergergerreg
dasda
adasdsad
test

B.2 Appendix B: Further Results

B.2.1 Survey Results

Age	(%)
18-25	0.25
25-35	0.33
35-45	0.20
45-55	0.17
+55	0.06
Gender	
Male	0.56
Female	0.43
Other	0.01
State	
Baden-Württemberg	0.10
Bavaria	0.17
Berlin	0.04
Brandenburg	0.04
Bremen	0.01
Hamburg	0.02
Hessen	0.08
Lower Saxony	0.11
Mecklenburg-Vorpommern	0.01
North Rhine-Westphalia	0.20
Rhineland-Palatinate	0.04
Saarland	0.03
Saxony	0.09
Saxony-Anhalt	0.02
Schleswig-Holstein	0.03
Thuringia	0.04
Education	
Less than high school	0.03
High school graduate	0.33
Bachelor's Degree	0.16
Graduate Degree	0.26
Vocational Qualification	0.23
Employment Status	
Paid Employee	0.47
Self-employed	0.21
Unemployed	0.08
Retired/Disabled	0.04
Student	0.21

Did vote in last elections?	
Yes	0.83
No	0.17
Interest in politics	
Not much interested	0.10
Somewhat interested	0.29
Interested	0.40
Very interested	0.21
Party	
CDU	0.12
SPD	0.15
AfD	0.12
FDP	0.13
Die Linke	0.09
Greens	0.15
CSU	0.04
Freie Wähler	0.02
PIRATEN	0.06
NPD	0.005
Other	0.14
Have you ever refrained from voicing an opinion on a political or social topic to avoid conflict?	
Often	0.13
Sometimes	0.65
Never	0.22
Are you more likely to hide your opinion from ...?	
Someone who disagrees with you	0.28
Someone whose opinion I don't know	0.72
N	199

Table A2: Demographic characteristics of the participants in the pre-experimental Survey.

B.2.2 Demographic Characteristics of the Participants

Table A3 shows the result of the post-experimental demographic questionnaire. A total of 2235 participants completed the questionnaire.

Table 2.2 shows the total percentage of donations by association (AfD and Pro Asyl) and experimental group (unexposed and exposed group) for all participants. unexposed group is the untreated condition in which no anti-immigrant comments were observed by the participants. exposed group, the treated group, features an increasing number of anti-immigrant comments with each consecutive forum page.

Two logistic models with donation as the dependent variable and different predictors were run to investigate which demographic characteristics have an effect of the donation decision to each of the organizations. Figure A3 shows the resulting coefficients for the model using data from the pre-experimental survey. Figure A2 shows the resulting coefficients for the model, using data from the experiment. Gender, employment status, education level, age, West Germany and size of the residency are used as predictors. For the models with the data from the experiment, the images participants saw before being asked about the donation are also added to the model. The images did not have any effect in the donation decision.

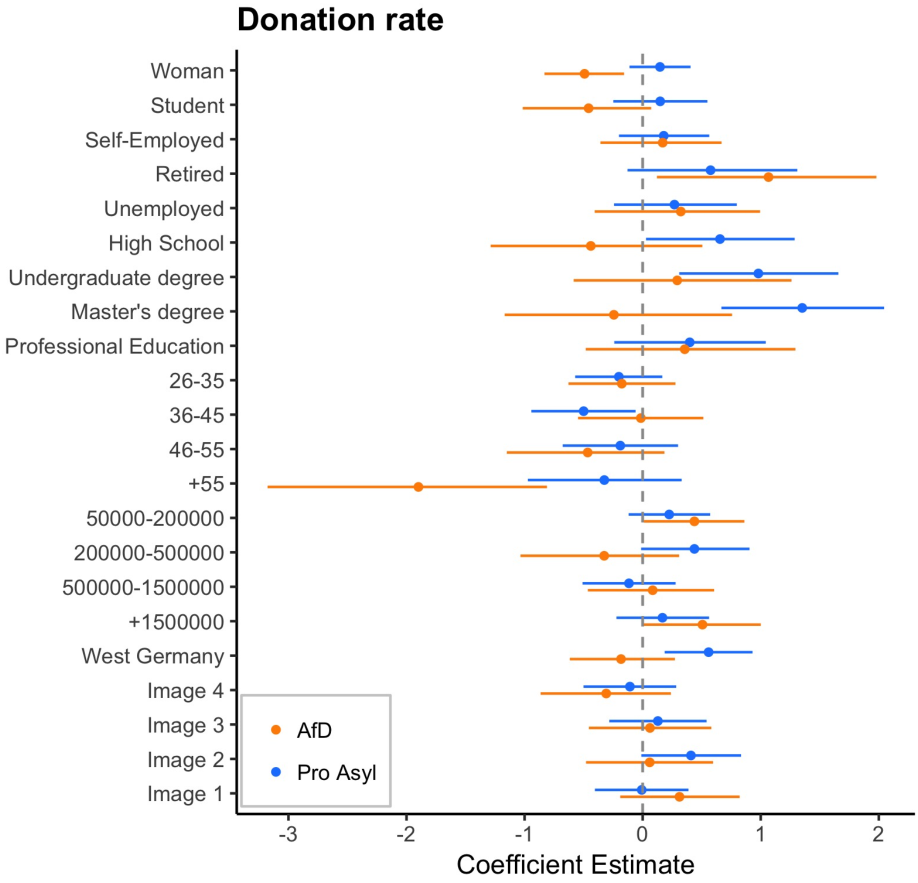


Figure A2: Coefficient Estimates of the Demographic Characteristics on Donation Rates for AfD and Pro Asyl.

Table A3: Characteristics of the participants based on the post-experiment questionnaire (N=2235).

	AfD, unexposed group	AfD, Exposed group	Pro Asyl, unexposed group	Pro Asyl, Exposed group	Total
	Freq.	Freq.	Freq.	Freq.	Freq. (%)
Gender					
Men	268	287	280	260	1095 (49%)
Women	288	284	271	248	1091 (49%)
Other	17	8	9	15	49 (2%)
Employment					
Unemployed	34	43	27	36	140 (6%)
Employed	332	298	325	270	1225 (55%)
Not active/retired	28	25	24	25	102 (5%)
Self-employed	79	88	71	81	319 (14%)
Student	100	125	113	111	449 (20%)
Education					
High School	210	201	195	170	776 (35%)
Bachelor	102	99	89	105	395 (18%)
Professional Qualification	152	143	164	139	598 (27%)
Master	92	106	89	88	375 (17%)
Primary or less	17	30	23	21	91 (4%)
East/West					
East	89	77	92	71	329 (15%)
West	484	502	468	452	1906 (85%)
Inhabitants					
<50.000	235	208	225	191	859 (38%)
50.000 – 200.000	125	137	125	108	495 (22%)
200.000 – 500.000	58	61	64	69	252 (11%)
500.000 – 1.500.000	70	88	83	67	308 (14%)
>1.500.000	85	85	63	88	321 (14%)
State					
Baden-Württemberg	75	73	90	60	298 (13%)
Bavaria	76	94	87	66	323 (15%)
Berlin	31	27	25	38	121 (5%)
Brandenburg	20	16	15	15	66 (3%)
Bremen	5	9	7	11	32 (1.4%)
Hamburg	20	18	14	25	77 (3.4%)
Hessen	50	49	33	39	171 (7.6%)
Mecklenburg-Vorpommern	8	8	9	7	32 (1.4%)
Lower Saxony	46	42	43	47	178 (8%)
North Rhine-Westphalia	135	135	126	121	517 (23%)
Rhineland-Palatinate	28	19	20	19	86 (4%)
Saarland	7	13	11	6	37 (1.7%)
Saxony	31	25	40	25	121 (5.4%)
Saxony-Anhalt	15	17	12	12	56 (2.4%)
Schleswig-Holstein	11	23	12	20	66 (3%)
Thuringia	15	11	16	12	54 (2.4%)
Age					
18-25	157	172	168	129	626 (28%)
26-35	217	193	185	200	795 (36%)
36-45	101	95	112	96	404 (18%)
46-55	66	77	57	64	264 (12%)
55+	32	42	38	34	146 (7%)
N					2235

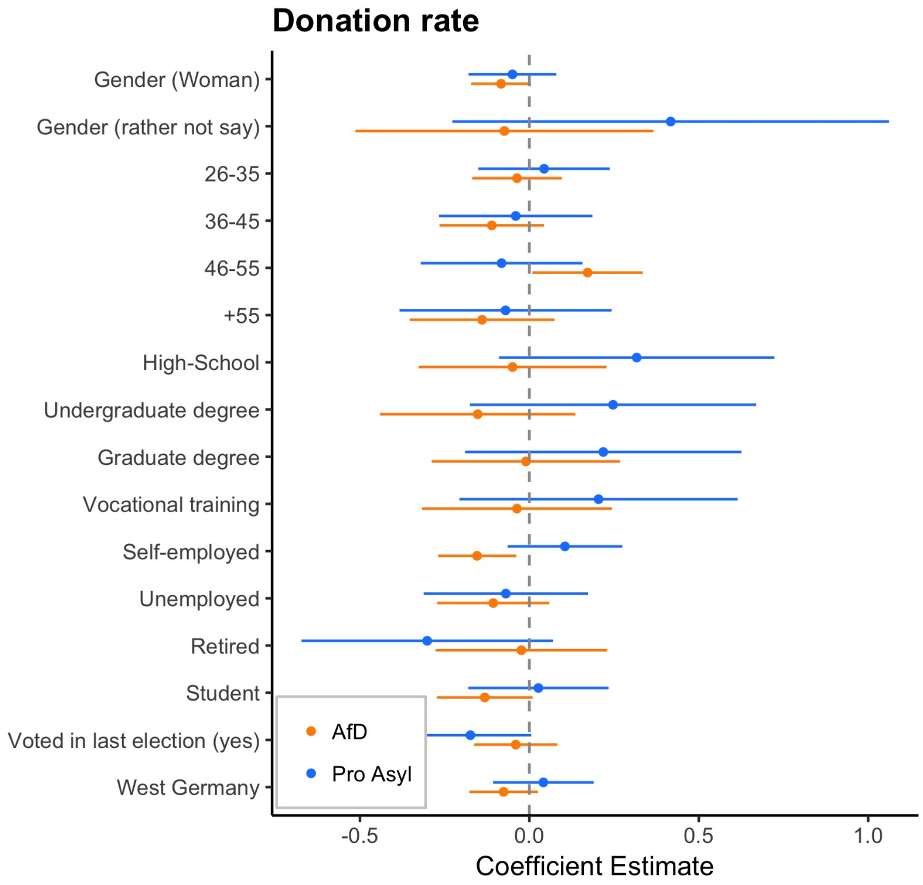


Figure A3: Coefficient Estimates of the Demographic Characteristics on Donation Rates for AfD and Pro Asyl from Pre-experimental Survey Data (199 participants).

B.2.3 Results by Association

In this section, the results of the experimental are presented. Results are presented with the output of the logistic models of donation as the dependent variable and treatment effects as predictors. The logistic models are presented along with the models' predictions computed as marginal effects (MEs), which shows the predicted values in their natural metric. The level of significance of the model predictions is also assessed. Because of the binary nature of the dependent variable, reporting the predicted levels can be more meaningful when interpreting the treatment effects, particularly the interaction terms (values contingent on values of other variables). The significance of the marginal effects is assessed using linear combination of the coefficients and tests for the first and second differences when necessary. Finally, the regression models with the polynomial predictors are shown.

Table A4: Logit regression estimates of the effect of the number of anti-immigrant comments observed before donation in the probability of donation. Models based on observations of the exposed group.

	(1) Pro Asyl All Obs.	(2) Pro Asyl Exposed	(3) AfD All obs.	(4) AfD Exposed
Dep. Var.: Donation				
1 Comment	0.0389	0.203	-0.211	-0.201
	(0.18)	(0.74)	(-0.75)	(-0.56)
3 Comments	0.0444	0.208	-0.122	-0.112
	(0.20)	(0.75)	(-0.43)	(-0.31)
6 Comments	0.203	0.367	0.100	0.110
	(0.94)	(1.33)	(0.38)	(0.32)
9 Comments	-0.0875	0.0766	-0.229	-0.219
	(-0.41)	(0.28)	(-0.75)	(-0.58)
Constant	0.477***	0.313	-1.609***	-1.619***
	(6.16)	(1.66)	(-15.89)	(-6.62)
Observations	1128	539	1155	574

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4 uses only observations from exposed group.

Table A5 shows the predicted probability of donation by organization and number of anti-immigrant comments observed before the donation decision. The predicted values based on model in Table A4. The first column shows the predicted values for each combination of number of comments and organization. The second column shows the difference to the baseline of 0 comments, that is, the predicted value for each number of comments is compared to having observed 0 comments, In the third column, the prediction for each given number of comments is compared to the prior number of comments. This way, the effect of observing three comment is compared to observing one comment, and observing one comments is compared to observing 0 comments.

In Table A6 the results from a logit model with donations as the dependent variable and Group (unexposed and exposed group), Time of donation (times 1 to 5), and their interaction as the predictors are shown (Group × Time). Columns 1 and 2 correspond to observations in the AfD conditions. Columns 3 and 4 show the results for Pro Asyl.

Table A7 and Table A8 show the predicted probabilities of donation for Pro Asyl and AfD, respectively, for the exposed group and time of donation. Table A7 and Table A8 show the predicted probability of donation by organization and number of anti-immigrant comments observed by the participants before the donation decision. The predicted values based on model in Table A6. The first column in each

Table A5: Predicted probability of donation by organization and number of anti-immigrant comments observed before the donation decision.

	Pr (Donation)	Difference to 0 comments (1st Difference)	Difference t-1 (1st Difference)
AfD 0 Comment(s)	0.17		
	(0.034)		
AfD 1 Comment(s)	0.14	-0.026	-0.026
	(0.031)	(0.046)	(0.046)
AfD 3 Comment(s)	0.15	-0.015	0.011
	(0.034)	(0.048)	(0.046)
AfD 6 Comment(s)	0.18	0.016	0.031
	(0.036)	(0.049)	(0.049)
AfD 9 Comment(s)	0.14	-0.028	-0.044
	(0.034)	(0.048)	(0.049)
Pro Asyl 0 Comment(s)	0.58		
	(0.045)		
Pro Asyl 1 Comment(s)	0.63	0.049	0.049
	(0.047)	(0.066)	(0.066)
Pro Asyl 3 Comment(s)	0.63	0.050	0.01
	(0.048)	(0.066)	(0.067)
Pro Asyl 6 Comment(s)	0.66	0.086	0.036
	(0.045)	(0.064)	(0.066)
Pro Asyl 9 Comment(s)	0.60	0.019	-0.067
	(0.048)	(0.066)	(0.066)

Notes: Standard errors in parentheses. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test.

table shows the predicted values for each combination of group and time. The second column shows the difference of each combination of group and time to donations in time one in their group. Column three shows the second differences of the comparison of each time to its baseline. It measures whether there are significant differences between the groups in the difference of each time to baseline time (Group × Time). The fourth column shows the difference of each time compared to time-1 (i.e., each time compared to the immediately preceding time). Column five shows the second differences of this difference. Finally, column 6 shows the difference between each time in each group. For example, the predicted value of Exposed group at Time 2 is compared to the predicted value of unexposed group at Time 2.

Table A6: Logit regression estimates of the effect of experimental group and treatment (donation time) in the probability of donation.

	(1) AfD Exposed	(2) AfD All obs.	(3) Pro Asyl Exposed	(4) Pro Asyl All obs.
Dep. Var.: Donation				
Time 2	-0.201	-0.470	0.203	0.236
	(-0.56)	(-1.24)	(0.74)	(0.88)
Time 3	-0.112	-0.0929	0.208	0.264
	(-0.31)	(-0.26)	(0.75)	(1.00)
Time 4	0.110	0.0302	0.367	-0.0535
	(0.32)	(0.09)	(1.33)	(-0.20)
Time 5	-0.219	0.249	0.0766	0.324
	(-0.58)	(0.75)	(0.28)	(1.19)
Exposed group		-0.0488		-0.0438
		(-0.14)		(-0.17)
Exposed group × Time 2		0.268		-0.0334
		(0.51)		(-0.09)
Exposed group × Time 3		-0.0188		-0.0557
		(-0.04)		(-0.14)
Exposed group × Time 4		0.0799		0.420
		(0.17)		(1.10)
Exposed group × Time 5		-0.468		-0.247
		(-0.93)		(-0.64)
Constant	-1.619***	-1.571***	0.313	0.357
	(-6.62)	(-6.55)	(1.66)	(1.91)
Observations	574	1155	539	1128

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A7: Predicted probability of donation to Pro Asyl by unexposed groupnd Time of donation decision.

	Pr(Donation)	Difference to T1	2nd Difference	Difference t-1	2nd Difference	Difference groups
Unexposed Time 1	0.59					
	(0.045)					
Unexposed Time 2	0.64	0.056		0.056		
	(0.044)	(0.063)		(0.063)		
Unexposed Time 3	0.65	0.062		0.006		
	(0.043)	(0.062)		(0.062)		
Unexposed Time 4	0.575	-0.013		-0.075		
	(0.046)	(0.065)		(0.063)		
Unexposed Time 5	0.66	0.076		0.089		
	(0.044)	(0.063)		(0.064)		
Exposed Time 1	0.58					-0.011
	(0.046)					(0.064)
Exposed Time 2	0.63	0.049	0.007	0.049	0.007	-0.018
	(0.046)	(0.066)	(0.091)	(0.066)	(0.091)	(0.064)
Exposed Time 3	0.63	0.050	0.012	0.050	0.005	-0.023
	(0.048)	(0.066)	(0.091)	(0.067)	(0.091)	(0.064)
Exposed Time 4	0.66	0.086	-0.099	0.086	-0.111	0.088
	(0.045)	(0.064)	(0.091)	(0.066)	(0.091)	(0.065)
Exposed Time 5	0.60	0.019	0.057	0.019	0.156	-0.068
	(0.048)	(0.066)	(0.092)	(0.066)	(0.092)	(0.065)

Notes: Standard errors in parentheses. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test.

Table A8: Probability of donation to AfD by group and time of donation decision.

	Pr(Donation)	Difference to T1	2nd Difference	Difference t-1	2nd Difference	Difference groups
Unexposed Time 1	0.17					
	(0.034)					
Unexposed Time 2	0.115	-0.057		-0.057		
	(0.031)	(0.045)		(0.045)		
Unexposed Time 3	0.16	-0.013		0.044		
	(0.034)	(0.049)		(0.046)		
Unexposed Time 4	0.18	0.004		0.017		
	(0.038)	(0.049)		(0.049)		
Unexposed Time 5	0.21	-0.038		0.034		
	(0.038)	(0.051)		(0.052)		
Exposed Time 1	0.17					-0.007
	(0.035)					(0.048)
Exposed Time 2	0.614	-0.026	-0.031	-0.026	-0.031	0.024
	(0.031)	(0.046)	(0.065)	(0.046)	(0.065)	(0.043)
Exposed Time 3	0.15	-0.015	0.002	0.011	0.033	-0.009
	(0.034)	(0.048)	(0.068)	(0.046)	(0.065)	(0.048)
Exposed Time 4	0.18	0.016	-0.011	0.031	-0.013	-0.067
	(0.036)	(0.049)	(0.069)	(0.049)	(0.069)	(0.066)
Exposed Time 5	0.14	-0.028	0.066	-0.044	0.078	-0.073
	(0.034)	(0.048)	(0.070)	(0.049)	(0.072)	0.051

Notes: Standard errors in parentheses. Significance levels: $^{***}p < 0.000$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test.

B.2.4 Results by Gender

In this section, the results are presented by gender. Only donations to the AfD are analyzed. Table A9 shows the results of a model with donations to AfD as the dependent variable and the number of comments participant observed before the donation decision as the predictors. Only observations from the are used in the models. The first column shows the effect of the number of observed comments for both men and women. Columns 2 and 3 show the same model for men and women. Finally, column 4 uses all observations and adds an interaction term (Comments Observed × Women).

Table A9: Logit model with donations to AfD as the dependent variable and the number of xenophobic comments participants observed before the donation decision as the predictors (only observations from the exposed group).

	(1)	(2)	(3)	(4)
	All Obs	Men	Women	All obs.
Dep. Variable: Donation				
1 Comments	-0.201	0.336	-0.882	0.336
	(-0.56)	(0.65)	(-1.64)	(0.65)
3 Comments	-0.112	0.344	-0.941	0.344
	(-0.31)	(0.69)	(-1.53)	(0.69)
6 Comments	0.110	0.725	-0.523	0.725
	(0.32)	(1.45)	(-1.05)	(1.45)
9 Comments	-0.219	0.257	-1.126 †	0.257
	(-0.58)	(0.51)	(-1.65)	(0.51)
Women				0.384
				(0.77)
Women × 1 Comment				-1.218
				(-1.63)
Women × 3 Comments				-1.285
				(-1.62)
Women × 6 Comments				-1.248 †
				(-1.77)
Women × 9 Comments				-1.383
				(-1.63)
Constant	-1.619***	-1.771***	-1.386***	-1.771***
	(-6.62)	(-4.63)	(-4.30)	(-4.63)
Observations	574	280	271	551

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A10: Logit model with donations to the AfD as the dependent variable and group (exposed and unexposed groups), Time of donation (times 1 to 5), Gender, and their interactions as the predictors.

	(1) All Observations	(2) Men	(3) Women	(4) All Observations
Dep. Variable: Donation				
Exposed group	-0.0488	-0.649	0.754	-0.649
	(-0.14)	(-1.32)	(1.40)	(-1.32)
Time 2	-0.470	-0.670	0.0198	-0.670
	(-1.24)	(-1.31)	(0.03)	(-1.31)
Time 3	-0.0929	-0.118	0.0970	-0.118
	(-0.26)	(-0.26)	(0.16)	(-0.26)
Time 4	0.0302	-0.632	0.754	-0.632
	(0.09)	(-1.33)	(1.38)	(-1.33)
Time 5	0.249	-0.194	0.876	-0.194
	(0.75)	(-0.42)	(1.64)	(-0.42)
Exposed group × Time 2	0.268	1.005	-0.902	1.005
	(0.51)	(1.38)	(-1.11)	(1.38)
Exposed group × Time 3	-0.0188	0.461	-1.038	0.461
	(-0.04)	(0.68)	(-1.22)	(0.68)
Exposed group × Time 4	0.0799	1.357*	-1.277†	1.357*
	(0.17)	(1.97)	(-1.73)	(1.97)
Exposed group × Time 5	-0.468	0.450	-2.002*	0.450
	(-0.93)	(0.66)	(-2.31)	(0.66)
Women				-1.018 †
				(-1.92)
Exposed group × Women				1.402 †
				(1.92)
Time 2 × Women				0.689
				(0.87)
Time 3 × Women				0.215
				(0.29)
Time 4 × Women				1.386†
				(1.91)
Time 5 × Women				1.070
				(1.52)
Exposed group × Time 2 × Women				-1.907†
				(-1.75)
Exposed group × Time 3 × Women				-1.499
				(-1.38)
Exposed group × Time 4 × Women				-2.634**
				(-2.60)
Exposed group × Time 5 × Women				-2.452*
				(-2.22)
Constant	-1.571***	-1.122***	-2.140***	-1.122***
	(-6.55)	(-3.65)	(-4.96)	(-3.65)
Observations	1155	548	559	1107

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In Table A10, the results from a logit model with donations to the AfD as the dependent variable and group (exposed and unexposed), Time of donation (times 1 to 5), Gender, and their interactions as the predictors are shown. Columns 1, 2, and 3 show a model with Group, Time, and the interaction term Group × Time as the predictor. The model is calculated for all observations (columnn 1), only men (column 2), and only women (column 3). Column 4 shows the results of a model with all observations and Group, Time, an interaction term for Group × Time, and an interaction term with 3 factors: Group × Time × Gender .

Table A11 and Table A12 show the predicted donation probabilities and the test for the first and second differences. All the values shown in these tables are computed based on Table A10.

Table A11 shows the results for men based on the second column in Table A10. The first column shows the predicted probability of donation to AfD for each donation time in each group. The second column shows the difference in predicted probability between each time and the baseline time (time 1) in its group. The third column shows the difference at each time between unexposed and exposed group. This table mirrors Table 2.5, which reports the results for women.

Table A12 shows the results for the interaction effects based on column 4 in Table A10. Column 1 shows the effects of the interaction between unexposed groupnd time for each gender. Similar to column 3 in Table 2.5. Column 2 shows the effects of the 3-way interaction Women × Group × Time, that is, whether the differences between groups at time X are different between women and men. In the last column, the differences between each time in Exposed group between men and women are presented.

Table A11: Probability of donation to AfD of men by experimental group and time of donation decision (treatments 1 to 5).

	Pr (Donation)	Difference to T1	Difference groups
Unexposed group Time 1	0.25		
	(0.057)		
Unexposed group Time 2	0.14	-0.103	
	(0.045)	(0.076)	
Unexposed group Time 3	0.22	-0.021	
	(0.06)	(0.082)	
Unexposed group Time 4	0.15	-0.098	
	(0.045)	(0.073)	
Unexposed group Time 5	0.21	-0.034	
	(0.056)	(0.080)	
Exposed group Time 1	0.15		-0.100
	(0.047)		(0.074)
Exposed group Time 2	0.19	0.047	0.049
	(0.055)	(0.072)	(0.074)
Exposed group Time 3	0.19	0.048	-0.031
	(0.05)	(0.069)	(0.078)
Exposed group Time 4	0.26	0.115	0.112
	(0.06)	(0.078)	(0.077)
Exposed group Time 5	0.18	0.035	-0.031
	(0.05)	(0.068)	(0.075)

Notes: Standard errors in parentheses. Significance levels: $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test.

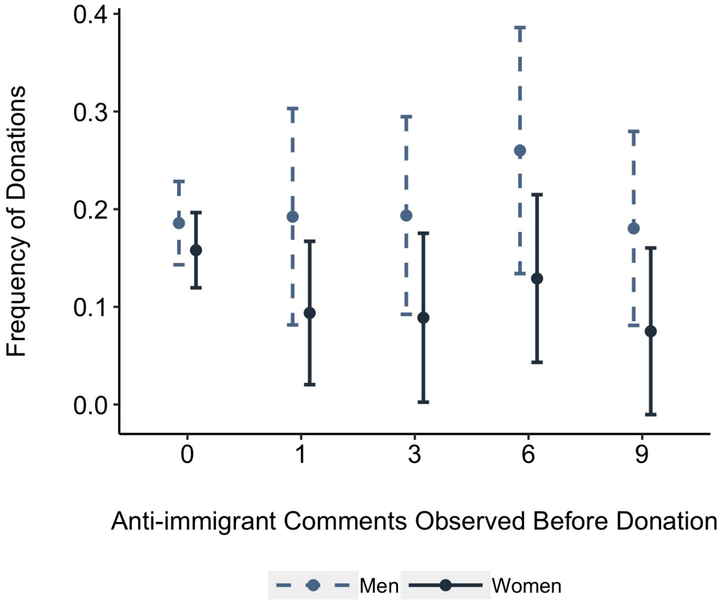


Figure A4: Frequency of donation in relation with the total number of racist comments seen before making the decision to donate to AfD. The plot uses all observations (N=2283). Error bars at the 95% confidence interval. The dashed line represents the results for men, and the solid line the results for women.

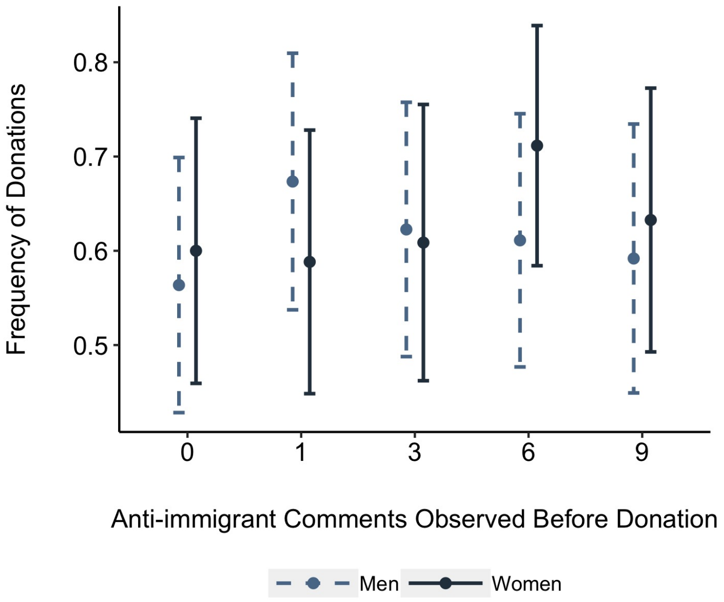


Figure A5: Frequency of donation in relation with the total number of racist comments seen before making the decision to donate to Pro Asyl (N= 1,128). Error bars at the 95% confidence interval. The dashed line represents the results for men, and the solid line the results for women.

Table A12: Difference in predicted probability of donation by gender for each Time and second differences of differences between times for men and women.

	Pr (Donation) (Group × Time)	2nd Differences (Women × Group × Time)	(Women × Time) Exposed group
Men Time 1	0.25		
	(0.057)		
Men Time 2	0.14		
	(0.045)		
Men Time 3	0.22		
	(0.06)		
Men Time 4	0.15		
	(0.045)		
Men Time 5	0.21		
	(0.056)		
Women Time 1	0.095	-0.095	0.055
	(0.066)	(0.066)	(0.07)
Women Time 2	-0.013	0.005	-0.099
	(0.055)	(0.098)	(0.066)
Women Time 3	-0.026	-0.005	-0.105
	(0.059)	(0.098)	(0.066)
Women Time 4	-0.071	0.183†	-0.131†
	(0.069)	(0.103)	(0.075)
Women Time 5	-0.145 *	0.114	-0.105
	(0.068)	(0.101)	(0.064)

Notes: Standard errors in parentheses. Significance levels: $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test.

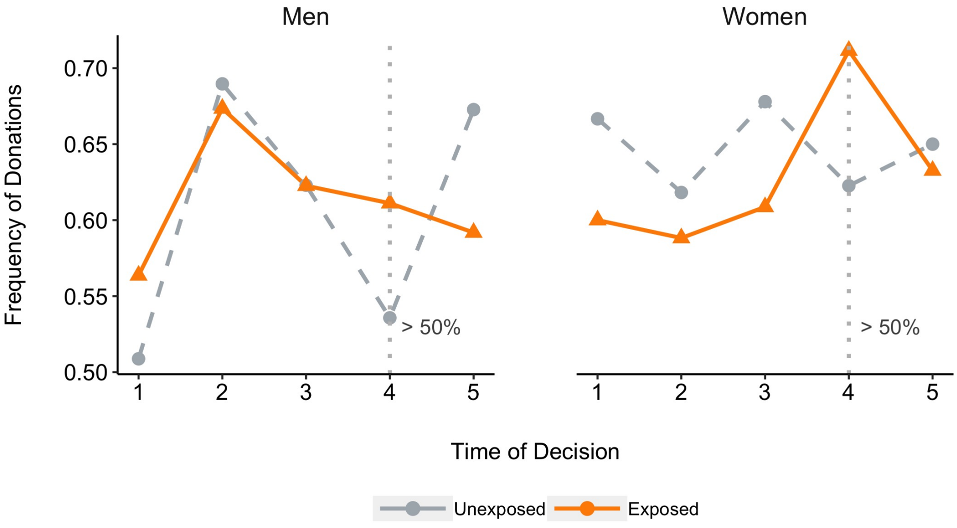


Figure A6: Frequency of donation to Pro Asyl in every stage of the forum for men and women making. The Orange line represents online forum pages were participants are exposed to an increasing number of anti-immigrant (Exposed group), compared to the unor unexposed group (dashed grey line). Only observations from Pro Asyl condition are used (Obs=1079).

Table A13: Logit model with donations to the AfD as the dependent variable and group (exposed and unexposed groups), Time of donation (times 1 to 5), Gender, and their interactions as the predictors.

	(1) All Observations	(2) Men	(3) Women	(4) All Observations
Dep. Variable: Donation				
Exposed group	-0.0438	0.221	-0.288	0.221
	(-0.17)	(0.58)	(-0.71)	(0.58)
Time 2	0.236	0.763*	-0.211	0.763*
	(0.88)	(1.97)	(-0.54)	(1.97)
Time 3	0.264	0.467	0.0513	0.467
	(1.00)	(1.25)	(0.13)	(1.25)
Time 4	-0.0535	0.108	-0.192	0.108
	(-0.20)	(0.29)	(-0.48)	(0.29)
Time 5	0.324	0.685	-0.0741	0.685
	(1.19)	(1.75)	(-0.19)	(1.75)
Exposed group × Time 2	-0.0334	-0.295	0.163	-0.295
	(-0.09)	(-0.52)	(0.29)	(-0.52)
Exposed group × Time 3	-0.0557	-0.222	-0.0149	-0.222
	(-0.14)	(-0.41)	(-0.03)	(-0.41)
Exposed group × Time 4	0.420	0.0880	0.690	0.0880
	(1.10)	(0.16)	(1.19)	(0.16)
Exposed group × Time 5	-0.247	-0.570	0.212	-0.570
	(-0.64)	(-1.02)	(0.37)	(-1.02)
Women				0.658
				(1.70)
Women × Exposed group 5				-0.509
				(-0.92)
Time 2 Women				-0.975
				(-1.76)
Time 3 Women				-0.416
				(-0.76)
Time 4 Women				-0.300
				(-0.55)
Time 5 Women				-0.760
				(-1.38)
Exposed group × Time 2 × Women				0.458
				(0.57)
Exposed group × Time 3 × Women				0.207
				(0.26)
Exposed group × Time 4 × Women				0.602
				(0.76)
Exposed group × Time 5 × Women				0.782
				(0.98)
Constant	0.357	0.0351	0.693*	0.0351
	(1.91)	(0.13)	(2.47)	(0.13)
N	1128	547	532	1079

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B.2.5 Results of the Polynomial Regression Models

	(1) AfD	(2) AfD	(3) Pro Asyl	(4) Pro Asyl
Dep. variable: Donations				
Constant	0.16 (0.03)**	0.31 (0.09)**	0.59 (0.04)**	0.47 (0.13)**
Comments	−0.03 (0.04)	−0.06 (0.04)	0.02 (0.06)	0.01 (0.06)
Comments 2	0.01 (0.01)	0.02 (0.01)$^{\cdot}$	−0.00 (0.02)	0.00 (0.02)
Comments 3	−0.00 (0.00)	−0.00 (0.00)$^{\cdot}$	−0.00 (0.00)	−0.00 (0.00)
Demographic Characteristics	*No*	*Yes*	*No*	*Yes*
R^2	0.00	0.04	0.00	0.06
Adj. R^2	-0.00	0.02	-0.00	0.04
Num. obs.	574	551	539	508

$^{***}p < 0$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$.

Table A14: Regression Estimates of donations to AfD and donations to Pro Asyl with a linear, quadratic and cubic terms of the number of anti-immigrant comments.

	(1) Women	(2) Women	(3) Men	(4) Men
Dep. variable: Donations				
Constant	0.19 (0.04)**	0.36 (0.13)**	0.16 (0.05)**	0.21 (0.13)†
Comments	−0.11 (0.05)†	−0.11 (0.05)*	0.01 (0.06)	−0.02 (0.06)
Comments 2	0.03 (0.01)†	0.03 (0.02)†	0.01 (0.02)	0.01 (0.02)
Comments 3	−0.00 (0.00)†	−0.00 (0.00)†	−0.00 (0.00)	−0.00 (0.00)
Demographic Characteristics	*No*	*Yes*	*No*	*Yes*
R^2	0.02	0.04	0.01	0.06
Adj. R^2	0.01	-0.00	-0.00	0.02
Num. obs.	271	271	280	280

$^{***}p < 0$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$

Table A15: Regression Estimates of donations to AfD of men and women with a linear, quadratic and cubic terms of the number of anti-immigrant comments.

	(1) Women	(2) Women	(3) Men	(4) Men
Dep. variable: Donations				
Constant	0.60 (0.06)**	0.61 (0.20)**	0.58 (0.06)**	0.40 (0.17)*
Comments	−0.04 (0.08)	−0.04 (0.09)	0.06 (0.08)	0.06 (0.08)
Comments 2	0.02 (0.02)	0.02 (0.02)	−0.02 (0.02)	−0.02 (0.02)
Comments 3	−0.00 (0.00)	−0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Demographic Characteristics	*No*	*Yes*	*No*	*Yes*
R^2	0.01	0.07	0.00	0.08
Adj. R^2	-0.00	0.02	-0.01	0.03
Num. obs.	248	248	260	260

$^{***}p < 0$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$

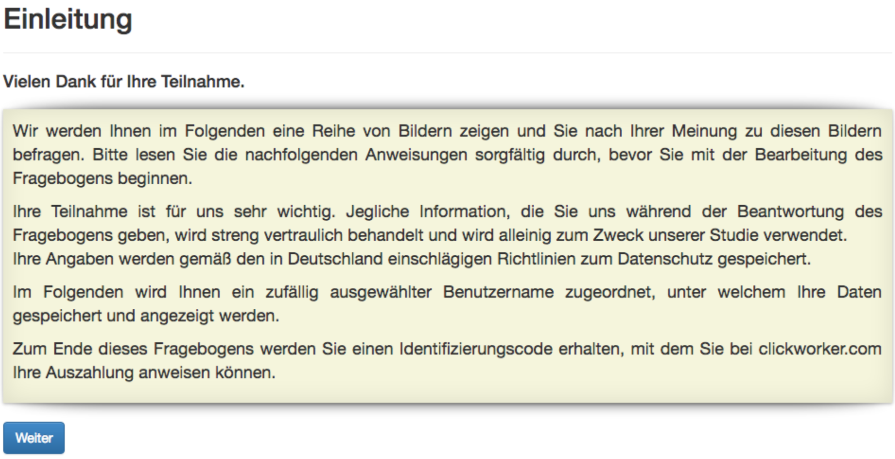
Table A16: Regression Estimates of donations to Pro Asyl of men and women with a linear, quadratic and cubic terms of the number of anti-immigrant comments.

Appendix C

Appendix Chapter 3

C.1 Appendix A: Materials

C.1.1 Instructions of the Experiment



Einleitung

Vielen Dank für Ihre Teilnahme.

Wir werden Ihnen im Folgenden eine Reihe von Bildern zeigen und Sie nach Ihrer Meinung zu diesen Bildern befragen. Bitte lesen Sie die nachfolgenden Anweisungen sorgfältig durch, bevor Sie mit der Bearbeitung des Fragebogens beginnen.

Ihre Teilnahme ist für uns sehr wichtig. Jegliche Information, die Sie uns während der Beantwortung des Fragebogens geben, wird streng vertraulich behandelt und wird alleinig zum Zweck unserer Studie verwendet. Ihre Angaben werden gemäß den in Deutschland einschlägigen Richtlinien zum Datenschutz gespeichert.

Im Folgenden wird Ihnen ein zufällig ausgewählter Benutzername zugeordnet, unter welchem Ihre Daten gespeichert und angezeigt werden.

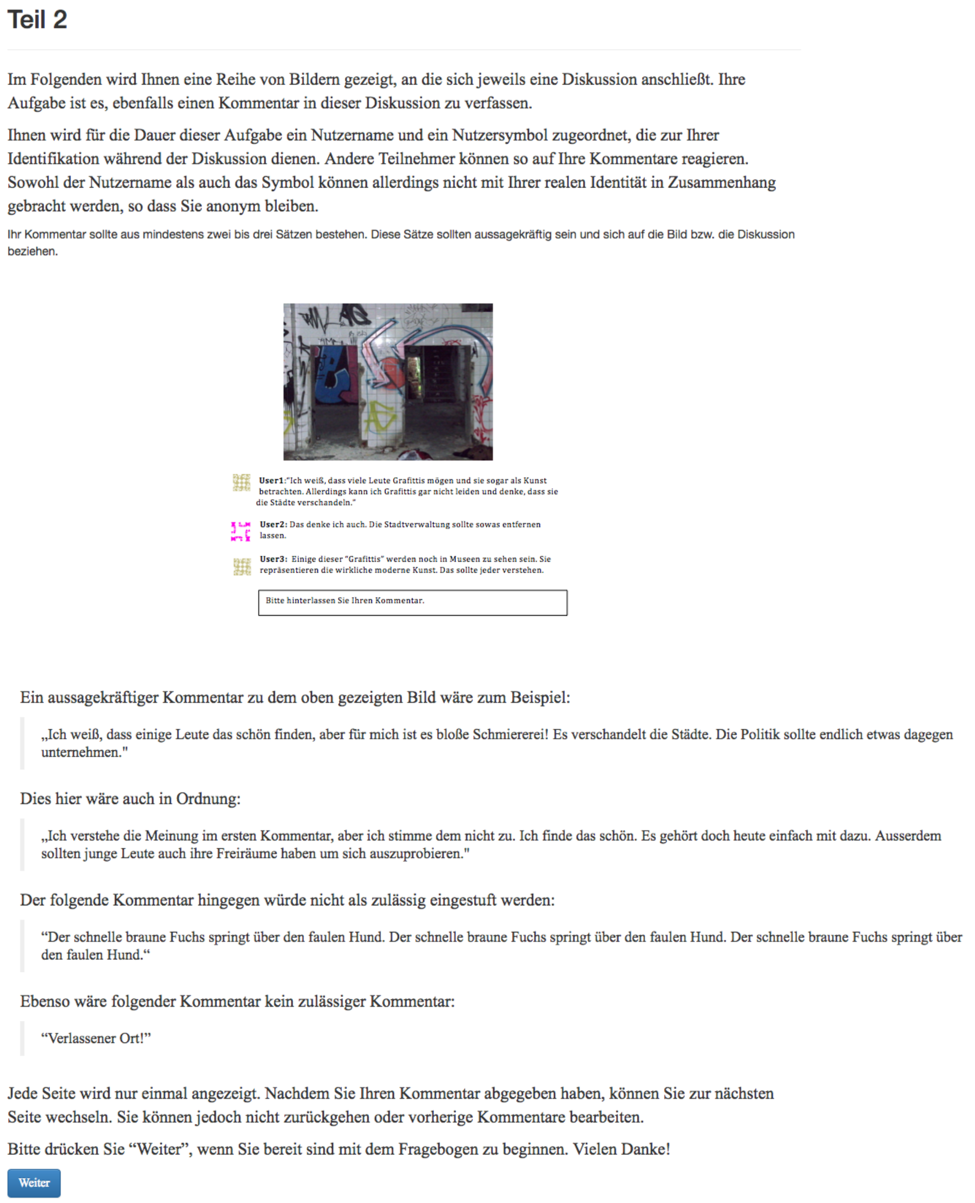
Zum Ende dieses Fragebogens werden Sie einen Identifizierungscode erhalten, mit dem Sie bei clickworker.com Ihre Auszahlung anweisen können.

Figure A1: Screenshot of the introduction page of the experiment in German. A translation of the instructions can be found below.

Introduction (In English)

Thank you for your participation.

We will show you a series of pictures and ask you comment on them. Please read the following instructions carefully before you begin the task. Your participation is very important to us. Any information you provide to us during the task will be strictly confidential and will be used solely for the purpose of our study. Your data will be stored in accordance with the relevant data protection guidelines in Germany. You will be assigned a random user name, and your input will be stored and displayed under this username. At the end, you will be given an identification code, which will allow you to claim your payment at `clickworker.com`



Teil 2

Im Folgenden wird Ihnen eine Reihe von Bildern gezeigt, an die sich jeweils eine Diskussion anschließt. Ihre Aufgabe ist es, ebenfalls einen Kommentar in dieser Diskussion zu verfassen.

Ihnen wird für die Dauer dieser Aufgabe ein Nutzername und ein Nutzersymbol zugeordnet, die zur Ihrer Identifikation während der Diskussion dienen. Andere Teilnehmer können so auf Ihre Kommentare reagieren. Sowohl der Nutzername als auch das Symbol können allerdings nicht mit Ihrer realen Identität in Zusammenhang gebracht werden, so dass Sie anonym bleiben.

Ihr Kommentar sollte aus mindestens zwei bis drei Sätzen bestehen. Diese Sätze sollten aussagekräftig sein und sich auf die Bild bzw. die Diskussion beziehen.

User1:"Ich weiß, dass viele Leute Grafittis mögen und sie sogar als Kunst betrachten. Allerdings kann ich Grafittis gar nicht leiden und denke, dass sie die Städte verschandeln."

User2: Das denke ich auch. Die Stadtverwaltung sollte sowas entfernen lassen.

User3: Einige dieser "Grafittis" werden noch in Museen zu sehen sein. Sie repräsentieren die wirkliche moderne Kunst. Das sollte jeder verstehen.

Bitte hinterlassen Sie Ihren Kommentar.

Ein aussagekräftiger Kommentar zu dem oben gezeigten Bild wäre zum Beispiel:

> „Ich weiß, dass einige Leute das schön finden, aber für mich ist es bloße Schmiererei! Es verschandelt die Städte. Die Politik sollte endlich etwas dagegen unternehmen."

Dies hier wäre auch in Ordnung:

> „Ich verstehe die Meinung im ersten Kommentar, aber ich stimme dem nicht zu. Ich finde das schön. Es gehört doch heute einfach mit dazu. Ausserdem sollten junge Leute auch ihre Freiräume haben um sich auszuprobieren."

Der folgende Kommentar hingegen würde nicht als zulässig eingestuft werden:

> "Der schnelle braune Fuchs springt über den faulen Hund. Der schnelle braune Fuchs springt über den faulen Hund. Der schnelle braune Fuchs springt über den faulen Hund.“

Ebenso wäre folgender Kommentar kein zulässiger Kommentar:

> "Verlassener Ort!"

Jede Seite wird nur einmal angezeigt. Nachdem Sie Ihren Kommentar abgegeben haben, können Sie zur nächsten Seite wechseln. Sie können jedoch nicht zurückgehen oder vorherige Kommentare bearbeiten.

Bitte drücken Sie "Weiter", wenn Sie bereit sind mit dem Fragebogen zu beginnen. Vielen Danke!

Figure A2: Screenshot of the instructions for the experiment in German. A translation of the instructions can be found below.

Instructions (Own Translation from the German original)

You will see a series of pictures with a discussion below. Your task is to join the discussion on the topic(s) depicted in the picture(s). Please write at least two to three sentences per discussion. These sentences should be meaningful and relate to the picture/discussion. A valid comment on the discussion above would be:

“I know that some people like them and even consider them to be art. However, I really dislike graffiti or “street art”, as some call it. I think it impoverishes the way a city looks.”

“I do understand the opinion in comment 1, although I pretty much disagree. Most of the places that are now covered by graffiti were previously abandoned and looked very dirty and ugly already.”

The following comment would not count as valid:

“The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.”

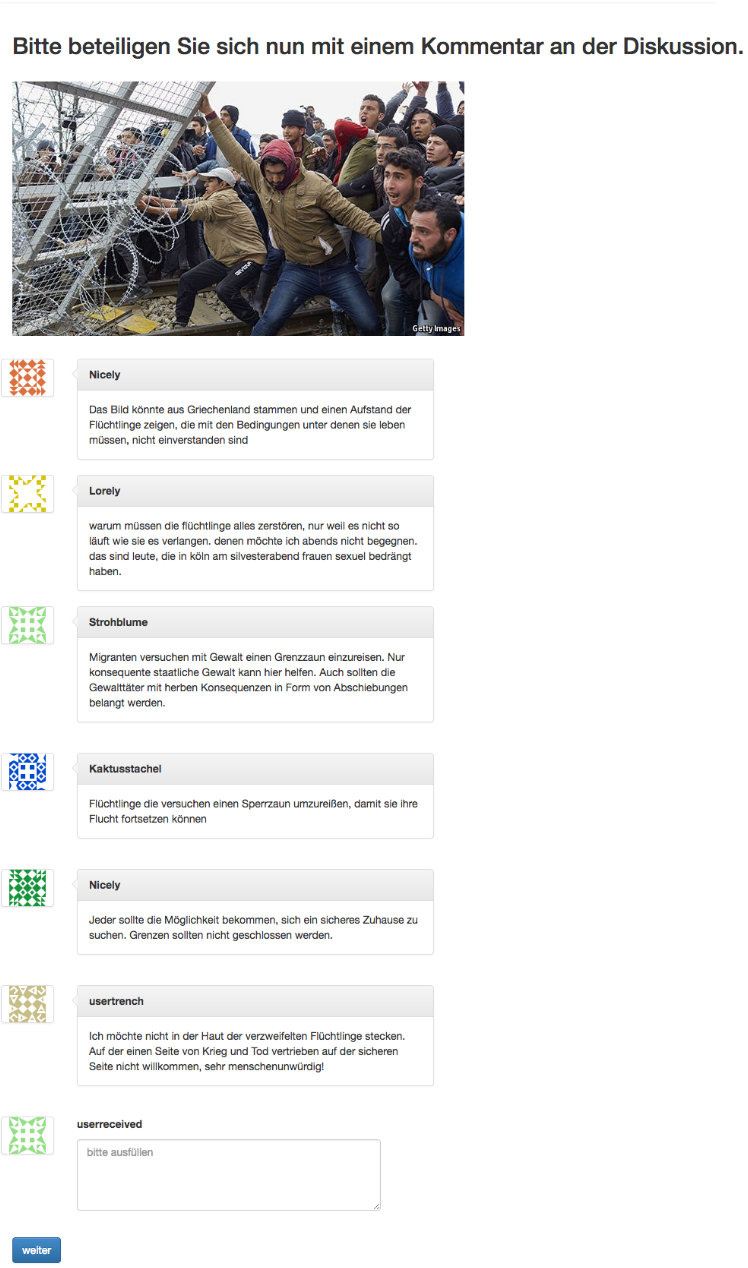
The following is not sufficient as a comment either:

“Abandoned place.”

Each page will be shown just once. Once you have finished with your comment you can go to the next page, but you cannot go back or edit previous comments. Please press the “start” button once you are ready to start the survey.

Thanks!

C.1.2 Screenshots of the Experimental Conditions of the Online Forum



Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.

Nicely

Das Bild könnte aus Griechenland stammen und einen Aufstand der Flüchtlinge zeigen, die mit den Bedingungen unter denen sie leben müssen, nicht einverstanden sind

Lorely

warum müssen die flüchtlinge alles zerstören, nur weil es nicht so läuft wie sie es verlangen. denen möchte ich abends nicht begegnen. das sind leute, die in köln am silvesterabend frauen sexuel bedrängt haben.

Strohblume

Migranten versuchen mit Gewalt einen Grenzzaun einzureisen. Nur konsequente staatliche Gewalt kann hier helfen. Auch sollten die Gewalttäter mit herben Konsequenzen in Form von Abschiebungen belangt werden.

Kaktusstachel

Flüchtlinge die versuchen einen Sperrzaun umzureißen, damit sie ihre Flucht fortsetzen können

Nicely

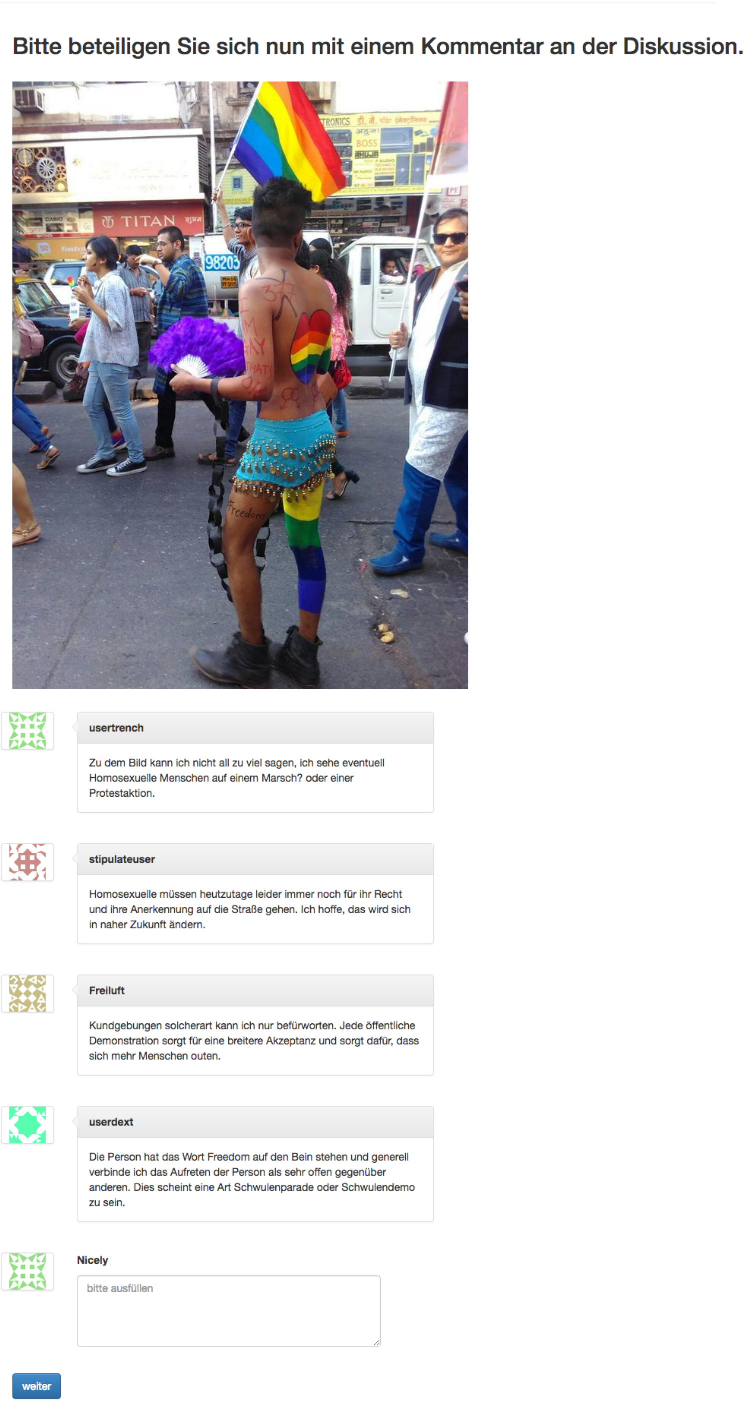
Jeder sollte die Möglichkeit bekommen, sich ein sicheres Zuhause zu suchen. Grenzen sollten nicht geschlossen werden.

usertrench

Ich möchte nicht in der Haut der verzweifelten Flüchtlinge stecken. Auf der einen Seite von Krieg und Tod vertrieben auf der sicheren Seite nicht willkommen, sehr menschenunwürdig!

userreceived

Figure A3: Screenshot of the discussion forum in the mixed condition in German. Examples of the comments in English can be found below.



Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.

usertrench

Zu dem Bild kann ich nicht all zu viel sagen, ich sehe eventuell Homosexuelle Menschen auf einem Marsch? oder einer Protestaktion.

stipulateuser

Homosexuelle müssen heutzutage leider immer noch für ihr Recht und ihre Anerkennung auf die Straße gehen. Ich hoffe, das wird sich in naher Zukunft ändern.

Freiluft

Kundgebungen solcherart kann ich nur befürworten. Jede öffentliche Demonstration sorgt für eine breitere Akzeptanz und sorgt dafür, dass sich mehr Menschen outen.

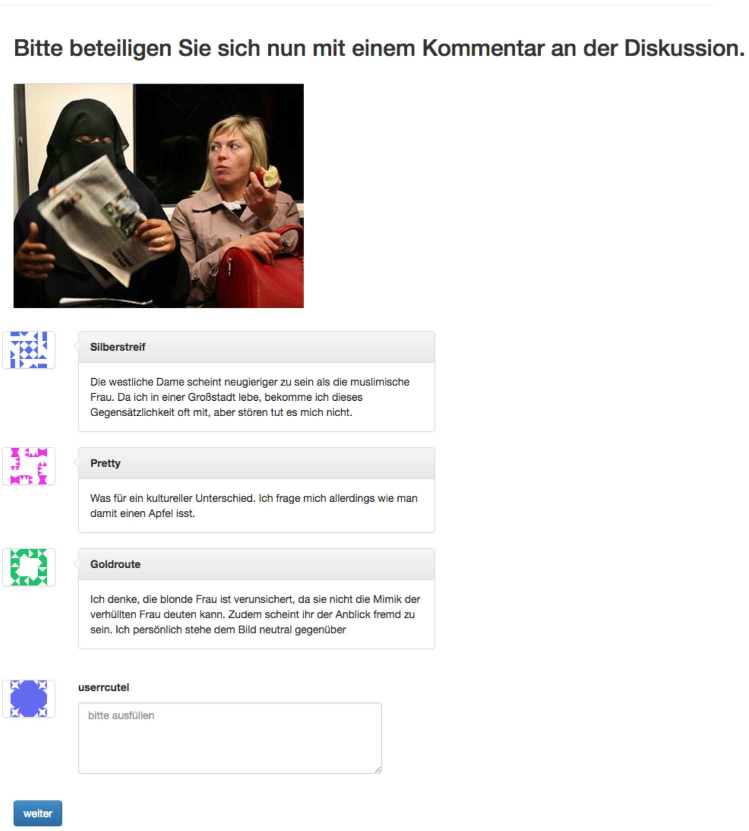
userdext

Die Person hat das Wort Freedom auf den Bein stehen und generell verbinde ich das Aufreten der Person als sehr offen gegenüber anderen. Dies scheint eine Art Schwulenparade oder Schwulendemo zu sein.

Nicely

bitte ausfüllen

Figure A4: Screenshot of the discussion forum in the neutral condition in German. Examples of the comments in English can be found below.



Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.

Silberstreif

Die westliche Dame scheint neugieriger zu sein als die muslimische Frau. Da ich in einer Großstadt lebe, bekomme ich dieses Gegensätzlichkeit oft mit, aber stören tut es mich nicht.

Pretty

Was für ein kultureller Unterschied. Ich frage mich allerdings wie man damit einen Apfel isst.

Goldroute

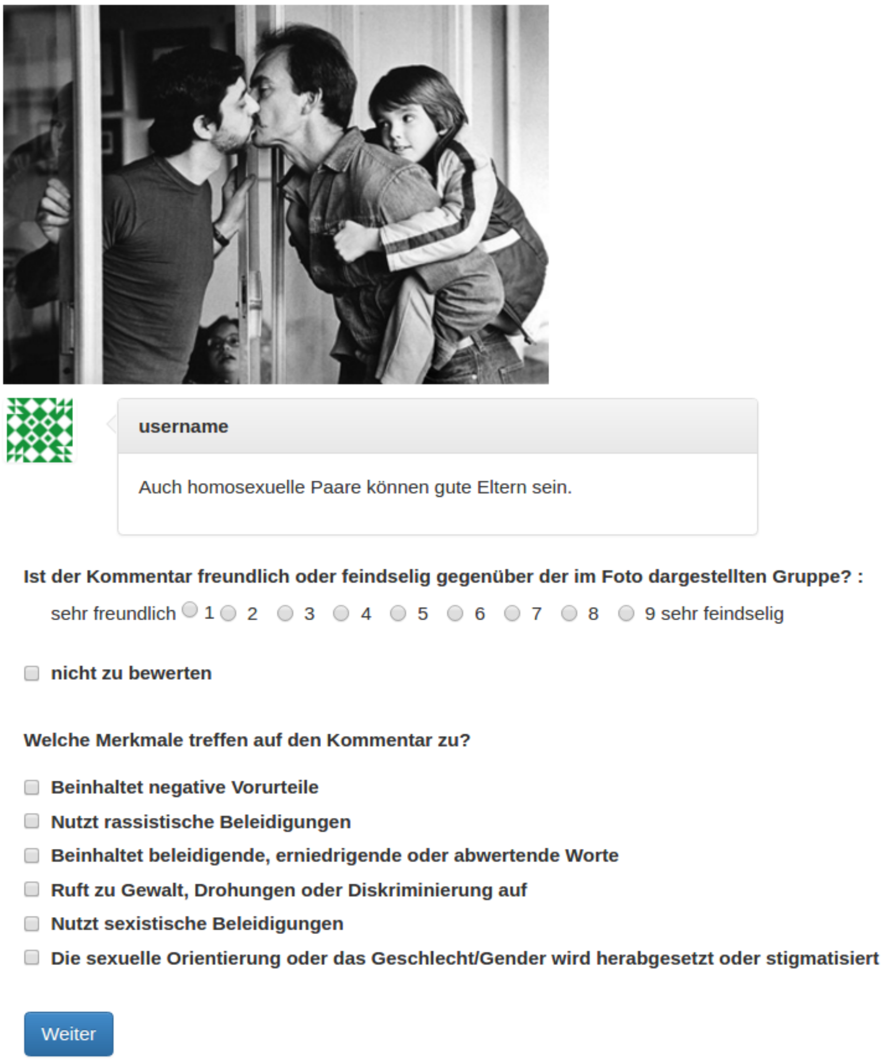
Ich denke, die blonde Frau ist verunsichert, da sie nicht die Mimik der verhüllten Frau deuten kann. Zudem scheint ihr der Anblick fremd zu sein. Ich persönlich stehe dem Bild neutral gegenüber

userrcutel

bitte ausfüllen

Figure A5: Screenshot of the discussion forum in the positive condition in German. Examples of the comments in English can be found below.

C.1.3 Screenshot of the Online Rating Form



username

Auch homosexuelle Paare können gute Eltern sein.

Ist der Kommentar freundlich oder feindselig gegenüber der im Foto dargestellten Gruppe? :

sehr freundlich ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 sehr feindselig

□ **nicht zu bewerten**

Welche Merkmale treffen auf den Kommentar zu?

□ **Beinhaltet negative Vorurteile**

□ **Nutzt rassistische Beleidigungen**

□ **Beinhaltet beleidigende, erniedrigende oder abwertende Worte**

□ **Ruft zu Gewalt, Drohungen oder Diskriminierung auf**

□ **Nutzt sexistische Beleidigungen**

□ **Die sexuelle Orientierung oder das Geschlecht/Gender wird herabgesetzt oder stigmatisiert**

Figure A6: Screenshot of the rating form used to rate the comments in German. A translation of the form can be found below.

(Translation of the rating form)

Is this comment friendly or hostile towards the group depicted in the picture? (from 1 to 9 where 1 is extremely friendly and 9 is extremely hostile)

□ It does not apply

Which items are found in the comments?

□ Contains negative stereotypes.

□ Uses racial slurs.

□ Contains words that are insulting, belittling, or diminishing.

□ Calls for violence, threat, or discrimination.

□ Uses sexual slurs.

□ Sexual orientation/gender used to ridicule or stigmatize.

C.1.4 Examples of Comments

Negative Comments

- □ Comment1029, mean hate score 8: “ich kann Schwule nicht ausstehen. die sollten sich psychiatrisch untersuchen lassen.” *(I cannot stand a gay people. They should have a psychiatric exam.)*
- □ Comment 391, mean hate score 8.33: “Randalierende Schmarotzer machen alles kaputt. Sie sollen wieder zurück gehen.” *(Rioting parasites destroy everything. They should go back again.)*
- □ Comment 1034,mean hate score 8.66: “Ekelhafte Zurschaustellung der eigenen Perversität. Ich hasse das.” *(Disgusting display of one's own perversity. I hate that.)*
- □ Comment 1261, mean hate score 8.66: “Flüchtlingskrise. Die können weiter wandern weg von europa. das sind nicht nur kriegsflüchtlinge, zu 90 prozent sind es nur sozialschmarotzer die hier machen können was die wollen.”*(Refugee crisis. They can continue walking away from Europe. They are not just war refugees, 90 per cent are nothing but social parasites who can do whatever they want here.)*
- □ Comment 1029, mean hate score 9: “Schwule Kerle sind das Letzte, was ich dulden würde. Schon gar nicht in der Öffentlichkeit.” *(Faggots are the last thing I would tolerate, especially in public.)*
- □ Comment 70, mean hate score 9: “Mit brutaler Gewalt wird versucht, das Sozialparadies Deutschland zu erreichen. Möge der Zaun ewig halten!” *(Brutal violence tries to reach the social paradise of Germany. May the fence last forever!)*

Neutral Comments

- □ Comment 641, mean hate score 5: “Leider ist dieses Thema noch hoch verpönt, vor allen Dingen wird es sicher nicht der kleine Junge im Leben leicht haben.” *(Unfortunately, this topic is still highly frowned upon, above all it will certainly not be easy for the little boy in life.)*
- □ Comment 892, mean hate score 5: “Generell bin ich dagegen sich in der Öffentlichkeit wild zu küssen. Aber gegen einen Kuss habe ich nichts.”*(In general I am against passionate kissing in public. But I have nothing against a kiss.)*

Positive Comments

- □ Comment 1572, mean hate score 1.66: “Das ist wirklich eine wunderschöne Szene. Es sieht nach einer absolut glücken Familie aus. Wahrscheinlich sind sie glücklicher als so manches hetero Paar.”*(This is really a wonderful scene. It looks like an absolutely happy family. They are probably happier than many heterosexual couples.)*
- □ Comment 1043, mean hate score 2: “Ich finde es gut das gleichgeschlechtliche paare sich immer mehr trauen dies in der Öffentlichkeit zu zeigen. Dies dann auch noch mit einer Bayrischen Tracht zu machen, einfach grossartig.”*(I think it is good that same-sex couples are increasingly daring to show this in public. To do this in traditional Bavarian clothes, just great.)*
- □ Comment 1082, mean hate score 2: “Wirklich ein zuckersüsses Bild! Eine kleine Familie und das Kind sieht so glücklich aus!” *(Truly a sweet picture! A little family and the kid looks so happy!)*
- □ Comment 469, mean hate score 3: “Zwei Männer in vermutlich einer Lebenspartnerschaft mit einem Kind. Ich finde es toll. Auch finde ich es wichtig, dass wir unsere Kinder heute gleich so erziehen, dass das alles auch normal ist. Und nicht auf dem alten Gedanken bleiben, dass eine Familie nur aus Mama, Papa und Kind(er) besteht, sondern es auch gleichgeschlechtliche "Eltern" gibt.” *(Two men presumably in a civil partnership with a child. I like it. Also, I think it is important that we raise our children nowadays with the message that this is normal. And not stick to the odd view that a family consists only of mum, dad and children, but that there are also same-sex parents.)*

- □ Comment 120, mean hate score 2: “Es handelt sich auf dem Bild offensichtlich um ein Menschenmasse von Flüchtlingen, welche auf dem Weg nach Europa sind, ich habe vollstes Verständnis für diese Menschen. Diese Menschen können nichts für den Krieg, der in ihren Ländern inszeniert wird, sie haben alles verloren und wollen (über)leben. Und das ist ihr (Menschen)recht !!!” *(The picture obviously shows a crowd of refugees on their way to Europe, I fully understand these people. There people are not responsible for the war that is staged in their countries, they have lost everything and want to live (survive). And that is their (human) right!!!*

- □ Comment 257, mean hate score 3: “Ich sehe hier eine junge Transfrau. Schön, mutig und selbstbewusst. Sie ist klasse.” *(I see a young trans woman here. Nice, brave and self-confident. She is great.)*

C.2 Appendix B: Further Analyses

C.2.1 Analysis of the Ratings

Figure A7 below shows the ratings per comment. The number of ratings per comment ranges from 1 to 13, with most comments rated between 5 and 10 times each.

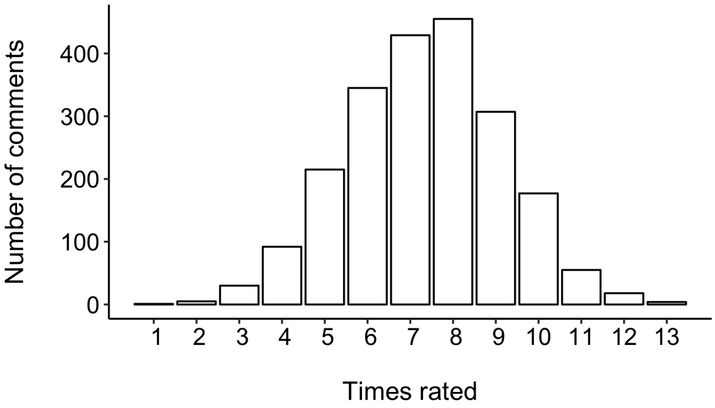


Figure A7: Number of ratings per comment.

Raters were asked to rate a total amount of 30 comments. The rate of attrition of the raters is relatively low, with only 13.7% of them abandoning the task before completion. Table A1 shows the percentage of raters that rated each number of times.

We then analyze the level of agreement of the ratings using both intraclass correlation (ICC), which measures similarity between the ratings of the same comment, and Spearman’s rank correlation test, which tests the statistical dependence between the rankings (i.e., the 1 to 9 scale). The ICC coefficient of the whole sample of ratings is 0.57, which is normally understood as a fair level of inter-rater agreement. We conducted robustness checks using a Spearman’s rank correlation in different subsets of the comments selected randomly from the whole dataset. The results are shown in Figure A8. The median Spearman’s rho coefficient is 0.48, with a maximum value of 0.52.[1]

[1] When comments with very few or too many ratings are not included in the analysis, the Spearman’s rho coefficient increases.

Number of ratings by rater	Freq.	Percent	Cum.
1	15	2.60	2.60
2	9	1.56	4.16
3	11	1.91	6.07
4	6	1.04	7.11
5	1	0.17	7.28
6	3	0.52	7.80
7	1	0.17	7.97
9	2	0.35	8.32
10	3	0.52	8.84
11	1	0.17	9.01
12	2	0.35	9.36
13	3	0.52	9.88
14	3	0.52	10.40
15	1	0.17	10.57
16	2	0.35	10.92
18	1	0.17	11.09
20	2	0.35	11.44
22	3	0.52	11.96
24	1	0.17	12.13
25	2	0.35	12.48
26	2	0.35	12.82
27	1	0.17	13.00
28	1	0.17	13.17
29	3	0.52	13.69
30	498	86.31	100.00
Total	577	100.00	

Table A1: Number of total comments rated per rater

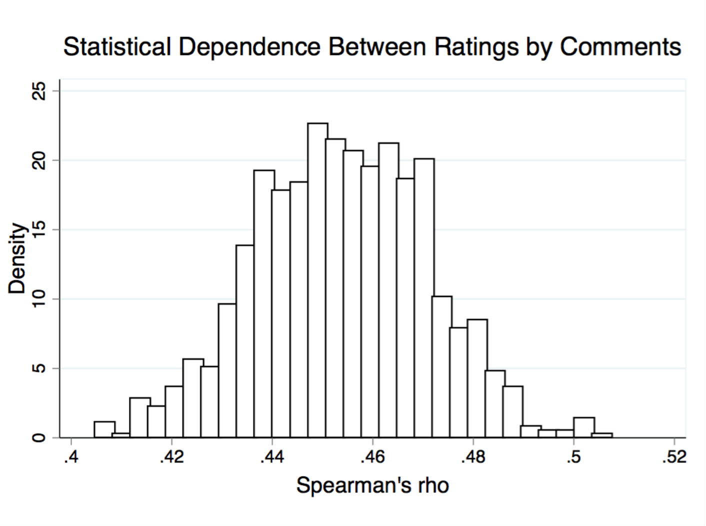


Figure A8: statistical dependence between the rankings.

C.2.2 Additional Information on our Identification Strategy

A key idea of the identification strategy is the assumption that the terrorist attacks had an effect of online public discussion on refugees and related topics, but did not have an effect on the topics we use for comparison,, i.e., LGBT and feminism. We can compare the number of searches on the different topics after the terrorist attacks. Figure A9 shows the search interest of the terms Transgender, Feminism, LGBT and related topics relative to the highest point on the chart for Germany between 1 June 2017 and 31 August, 2016. The interest could be read as follows: a value of 100 is the peak of popularity for the term; a value of 50 means that the term is half as popular; likewise, a score of 0 means the term was less than 1% as popular as the peak. The popularity of these terms was not affected by the terrorist attacks (grey dashed lines). This can also be compared to the data in Figure 3.4.

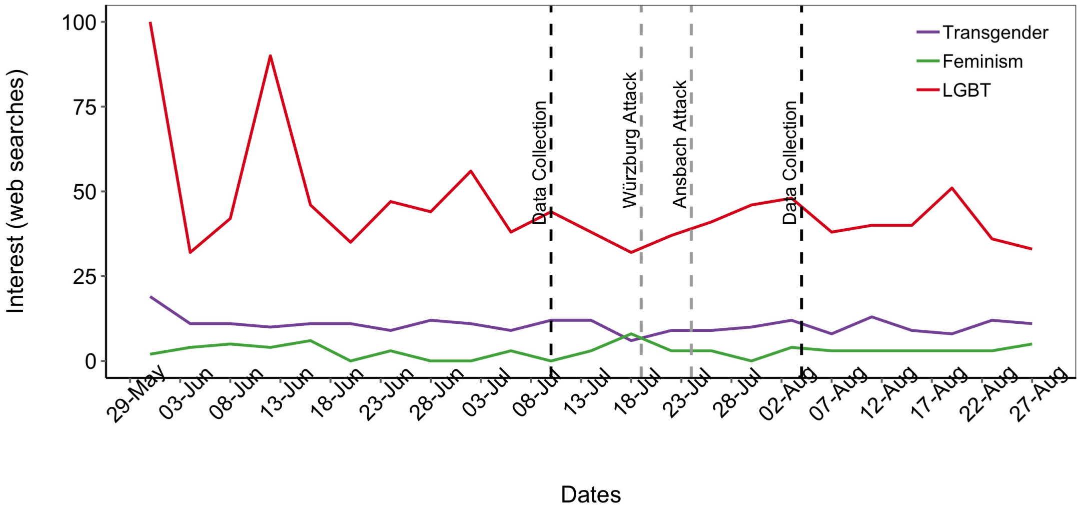


Figure A9: Relative interest (web searches) in Germany after the terrorist attacks. Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. The dashed lines represent the events of interest: dates of data collection and attacks.

C.2.3 Sociodemographic Characteristics of the Population from where the Sample was Recruited

Female	55%
Age	
18-24	28%
25-34	42%
35-44	17%
>45	13%
Employment status	
Student	29%
Employee	26%
Self-employed	15%
Other	20%
N.S	10%

Table A2: Sociodemographic characteristics of the population from where participants were recruited

C.2.4 Statistical Models

Our first statistical model is displayed in column 1 in Table 3.4 and is estimated only for the mixed condition using the comments on refugees:

$$Y_{ijk} = \beta_0 + \beta_1 AfterAttacks_{ijk} + u_i + v_j + \epsilon_{ijk} \tag{C.1}$$

where $u_i \sim N(0, \sigma_u)$ and $v_j \sim N(0, \sigma_j)$. The model estimates the rating Y_{ijk} for comment k by participant i on picture j. It is worth noting again that u_i (the random effect on the participant level) and v_j (the random effect on the picture level) are crossed random effects, that is pictures are neither fully nested in participants nor are participants fully nested in pictures. It means precisely that participants appear in more than one picture. This applies to all the following models except for the quantile regression.

The second model (Model 2 in Table 3.4) is specified as follows:

$$Y_{ijk} = \beta_0 + \beta_1 AfterAttacks_{ijk} + \beta_2 Topic_{ijk} + \beta_2 (AfterAttacks_{ijk} \times Topic_{ijk}) + u_i + v_j + \epsilon_{ijk} \tag{C.2}$$

It introduces an interaction term between topic (Refugees or Gender Rights) and the terrorist attack.

Model 3 in Table 3.4 estimates the effects of the terrorist attacks in each experimental condition on comments about refugees:

$$Y_{ijk} = \beta_0 + \beta_1 AfterAttacks_{ijk} + \beta_2 Condition_{ijk} + \beta_3 (AfterAttacks_{ijk} \times Condition_{ijk}) + u_i + v_j + \epsilon_{ijk} \tag{C.3}$$

Finally, the full model is displayed in column 4 of Table 3.4. It introduces three-way interactions between the attack, the topic, and the treatments and is estimated for the full data set:

$$Y_{ij} = \beta_0 + \beta_1 AfterAttacks_{ijk} + \beta_2 Topic_{ijk} + \beta_3 Condition_{ijk} + \beta_4 (AfterAttacks_{ijk} \times Condition_{ijk}) \\ + \beta_5 (AfterAttacks_{ijk} \times Condition_{ijk} \times Topic_{ijk}) + u_j + v_j + \epsilon_{ijk} \tag{C.4}$$

The quantile regression models reported in subsection 3.5.3 have the standard form:

$$Y_i = \beta_0 + \beta_1 (AfterAttacks_i \times Conditions_i) + u_i \tag{C.5}$$

The quantile regression estimator for quantile q minimizes the objective function:

$$Q(\beta_q) = \sum_{i:y_i \geqslant x_i \beta}^{N} q|y_i - x_i^{'}\beta_q| - \sum_{i:y_i \geqslant x_i \beta}^{N} (1-q)|y_i - x_i^{'}\beta_q| \tag{C.6}$$

We calculated quantile regression estimates for percentiles 10th to 95th in increments of 5.

C.2.5 Additional Results for the Quantile Regression

Figure A10 shows the coefficient of the effect of the terrorist attacks for the different topics and treatments for the quantiles 0.10 to 0.95 of the distribution of the hate score, as well as the density curves[2] before and after the terrorist attacks. The top left plot in Figure A10 shows the effect of the terror attack in the mixed treatment in Refugees. The terrorist attacks increased the average hate speech for all quantiles of the distribution (the mean increase is 0.56 points), but the effect is stronger in the most "hateful" quantiles of the distribution.

[2] A Gaussian kernel was used to estimate the density curves.

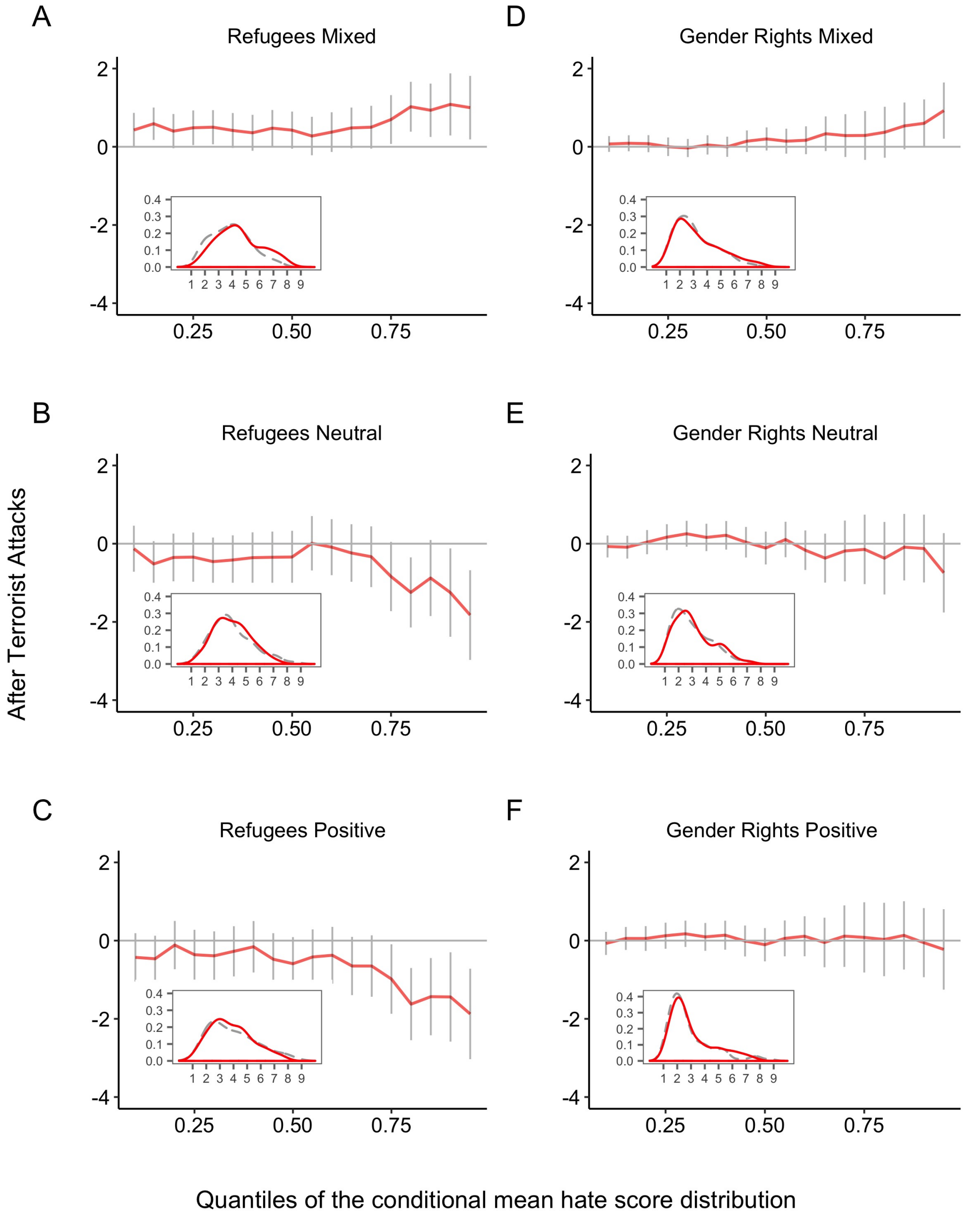


Figure A10: The plots depict the estimated model coefficients of the effect on terrorist attacks for quantiles 0.10 to 0.95 for each combination of topic and experimental condition. The grey vertical lines represent the confidence interval of the quantile regression coefficients for the effect of terrorist attacks. The plots also show the density distribution of the hate score before (dashed line) and after (solid line) the terrorist attacks. The left column shows the plots for Refugees. From top to bottom: mixed treatment (A), neutral treatment (B), and positive treatment (C). The column on the right shows the results for Gender Rights, from top to bottom: mixed treatment (D), neutral treatment (E), and positive treatment (F).

Table A3: Estimated regression coefficients for percentiles 10th to 95th of the distribution of the hate score of the treatment effect of the terrorist attacks in each experimental condition in comments on refugees. Confidence intervals at 95% Level.

Quantile	Intercept	Mixed Condition	Neutral Condition	Positive Condition
		Coef (up , lb)	Coef (up , lb)	Coef (up , lb)
0.10	1.571	0.429(−0.009 , 0.866)	−0.129 (−0.716 , 0.459)	−0.429 (−1.047 , 0.189)
0.15	1.536	0.589(0.176 , 1.002)	−0.518 (−1.101 , 0.065)	−0.464 (−1.053 , 0.124)
0.20	2.200	0.400(−0.037 , 0.837)	−0.352 (−0.962 , 0.257)	−0.114 (−0.732 , 0.503)
0.25	2.229	0.486(0.042 , 0.929)	−0.343 (−0.972 , 0.286)	−0.361 (−0.998 , 0.277)
0.30	2.500	0.500(0.063 , 0.937)	−0.458 (−1.074 , 0.157)	−0.389 (−1.014 , 0.237)
0.35	2.833	0.417(−0.029 , 0.863)	−0.417 (−1.047 , 0.213)	−0.274 (−0.915 , 0.368)
0.40	3.143	0.357(−0.102 , 0.816)	−0.357 (−1.004 , 0.290)	−0.157 (−0.815 , 0.50)
0.45	3.190	0.476(0.012 , 0.940)	−0.348 (−1.005 , 0.309)	−0.476 (−1.142 , 0.190)
0.50	3.484	0.425(−0.050 , 0.899)	−0.341 (−1.011 , 0.330)	−0.591 (−1.271 , 0.089)
0.55	3.850	0.275(−0.217 , 0.767)	0.011 (−0.684 , 0.706)	−0.418 (−1.123 , 0.287)
0.60	3.875	0.375(−0.132 , 0.882)	−0.089 (−0.805 , 0.626)	−0.375 (−1.102 , 0.352)
0.65	3.893	0.482(−0.038 , 1.002)	−0.238 (−0.973 , 0.498)	−0.649 (−1.394 , 0.096)
0.70	4.000	0.500(−0.047 , 1.047)	−0.333 (−1.110 , 0.443)	−0.650 (−1.435 , 0.135)
0.75	4.179	0.696(0.073 , 1.320)	−0.839 (−1.725 , 0.046)	−0.982 (−1.870 , −0.094)
0.80	4.119	1.024(0.387 , 1.661)	−1.246 (−2.142 , −0.350)	−1.624 (−2.547 , −0.701)
0.85	4.467	0.933(0.251 , 1.615)	−0.878 (−1.850 , 0.095)	−1.433 (−2.421 , −0.446)
0.90	4.667	1.083(0.289 , 1.877)	−1.250 (−2.380 , −0.120)	−1.440 (−2.582 , −0.299)
0.95	5.500	1.000(0.188 , 1.812)	−1.825 (−2.972 , −0.678)	−1.875 (−3.031 ,−0.719)

Table A4: Estimated regression coefficients for percentiles 10th to 95th of the distribution of the hate score of the treatment effect of the terrorist attacks in each experimental condition in comments on gender rights. Confidence intervals at 95% Level

Quantile	Intercept	Mixed Condition	Neutral Condition	Positive Condition
		Coef (up , lb)	Coef (up , lb)	Coef (up , lb)
0.10	1.429	0.071 (−0.130 , 0.273)	−0.071 (−0.352 , 0.209)	−0.071 (−0.366 , 0.223)
0.15	1.536	0.089 (−0.115 , 0.294)	−0.089 (−0.381 , 0.202)	0.055 (−0.238 , 0.348)
0.20	1.698	0.079 (−0.138 , 0.297)	0.040 (−0.270 , 0.349)	0.054 (−0.262 , 0.370)
0.25	2.000	0.00 (−0.237 , 0.237)	0.167 (−0.165 , 0.498)	0.125 (−0.208 , 0.458)
0.30	2.175	−0.032 (−0.263 , 0.200)	0.254 (−0.077 , 0.585)	0.175 (−0.162 , 0.511)
0.35	2.238	0.048 (−0.198 , 0.294)	0.161 (−0.189 , 0.510)	0.095 (−0.261 , 0.451)
0.40	2.429	0.00 (−0.258 , 0.258)	0.214 (−0.151 , 0.580)	0.139 (−0.232 , 0.510)
0.45	2.429	0.143 (−0.132 , 0.418)	0.038 (−0.353 , 0.428)	−0.009 (−0.405 , 0.386)
0.50	2.600	0.200 (−0.096 , 0.496)	−0.111 (−0.531 , 0.309)	−0.105 (−0.530 , 0.320)
0.55	2.714	0.143 (−0.176 , 0.462)	0.107 (−0.346 , 0.560)	0.057 (−0.402 , 0.516)
0.60	3.000	0.167 (−0.189 , 0.523)	−0.167 (−0.672 , 0.339)	0.111 (−0.401 , 0.623)
0.65	3.000	0.333 (− 0.109 , 0.776)	− 0.369 (−0.995 , 0.257)	−0.048 (−0.681 , 0.586)
0.70	3.429	0.286 (−0.261 , 0.833)	−0.186 (−0.962 , 0.591)	0.114 (−0.672 , 0.900)
0.75	3.822	0.289 (−0.335 , 0.913)	−0.146 (−1.034 , 0.742)	0.082 (−0.815 , 0.980)
0.80	4.057	0.371 (−0.280 , 1.023)	−0.371 (−1.298 , 0.555)	0.029 (−0.910 , 0.967)
0.85	4.267	0.533 (−0.065 , 1.131)	−0.089 (−0.939 , 0.761)	0.133 (−0.740 , 1.007)
0.90	4.800	0.600 (−0.011 , 1.211)	−0.124 (−0.991 , 0.743)	−0.052 (−0.935 , 0.830)
0.95	4.950	0.925 (0.207 , 1.643)	−0.744 (−1.759 , 0.270)	−0.225 (−1.253 , 0.803)

References

Abbott, A. (2007). Mechanisms and relations. *Sociologica*, *1*(2), 0–0.

Allport, G. W. (1979). *The nature of prejudice.* New York: Basic books.

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237.

Andersen, T. M., Bertola, G., Driffill, J., Fuest, C., James, H., Sturm, J.-E., & Uroševic, B. (2017). Immigration and the refugee crisis – Can Europe rise to the challenge? *EEAG Report on the European Economy*, 82–101.

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics*, *30*, 98–108.

Awan, I., & Zempi, I. (2017). 'I will blow your face OFF' – Virtual and physical world anti-Muslim hate crime. *The British Journal of Criminology*, *57*(2), 362–380.

Bader, F., & Keuschnigg, M. (2018). *Conducting large-scale online experiments on a crowdsourcing platform.* SAGE Publications Ltd.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817.

Berinsky, A. J. (1999). The two faces of public opinion. *American Journal of Political Science*, 1209–1230.

Betz, H.-G. (1994). *Radical right-wing populism in western europe.* Springer.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms.* Cambridge: Cambridge University Press.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms.* New York, NY: Oxford University Press.

Bicchieri, C., & Chavez, A. K. (2013). Norm manipulation, norm evasion: Experimental evidence. *Economics and Philosophy*, *29*(02), 175–198.

Bicchieri, C., & Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice, Forthcoming.*

Bicchieri, C., & Funcke, A. (2018). Norm change: Trendsetters and social structure. *Social Research: An International Quarterly*, *85*(1), 1–21.

Bicchieri, C., & McNally, P. (2018). Shrieking sirens: Schemata, scripts, and social norms. how change occurs. *Social Philosophy and Policy*, *35*(1), 23–53.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, *22*(2), 191–208.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Binns, A. (2012). Don't feed the trolls! managing troublemakers in magazines' online communities. *Journalism Practice*, *6*(4), 547–562.

Blanchard, F. A., Crandall, C. S., Brigham, J. C., & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, *79*(6), 993.

Blinder, S., Ford, R., & Ivarsflaten, E. (2013). The better angels of our nature: How the antiprejudice norm affects policy and party preferences in Great Britain and Germany. *American Journal of Political Science*, *57*(4), 841–857.

Boomgaarden, H. G., & de Vreese, C. H. (2007). Dramatic real-world events and public opinion dynamics: Media coverage and its impact on public reactions to an assassination. *International Journal of Public Opinion Research*, *19*(3), 354–366.

Burgoon, M., Alvaro, E., Grandpre, J., & Voulodakis, M. (2002). Revisiting the theory of psychological reactance. In *The persuasion handbook* (pp. 213–232). Thousand Oaks, CA: Sage.

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., . . . Voss, A. (2014). Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 206.

Bursztyn, L., Egorov, G., & Fiorin, S. (2017, May). *From extreme to mainstream: How social norms unravel.* (Available at NBER working paper series: `http://www.nber.org/papers/w23415`)

Byers, B. D., & Jones, J. A. (2007). The impact of the terrorist attacks of 9/11 on anti-Islamic hate crime. *Journal of Ethnicity in Criminal Justice*, *5*(1), 43–56.

Cantoni, D., Hagemeister, F., & Westcott, M. (2017). Persistence and activation of right-wing political ideology. *Available at Munich Discussion Paper No. 2017-14*.

Centola, D. (2018). *How behavior spreads: The science of complex contagions* (Vol. 3). Princeton University Press.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American journal of Sociology*, *113*(3), 702–734.

Centola, D., Willer, R., & Macy, M. (2005). The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, *110*(4), 1009–1040.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree - an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Cheng, J., Danescu-Niculescu-Mizil, C., Leskovec, J., & Bernstein, M. (2017). Anyone can become a troll. *American Scientist*, *105*(3), 152.

Cialdini, R. B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika*, *72*(2), 263.

Cialdini, R. B. (2009). *Influence: Science and practice* (Vol. 4). Pearson education Boston.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology* (Vol. 24, pp. 201–234). Elsevier.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 151–192). New York: McGraw-Hill.

Citron, D. K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, *91*, 1435.

Coffé, H. (2018). Gender and the radical right. In *The oxford handbook of the radical right.* Oxford University Press.

Coleman, J. S. (1990a). The emergence of norms. In M. Hechter, K.-D. Opp, & R. Wippler (Eds.), *Social institutions: Their emergence, maintenance, and effects* (pp. 35–39). New York: Aldine de Gruyter.

Coleman, J. S. (1990b). *Foundations of social theory.* Cambridge: Harvard University Press.

Coleman, J. S. (1990c). Norm-generating structures. *The limits of rationality*, 250–273.

Connolly, K. (2016, July). *Pressure grows on angela merkel to start closing germany's open door. The Guardian*, July 25. (Retrieved June, 2017 (https://www.theguardian.com/world/2016/jul/25/pressure-grows-on-angela-merkel-to-start-closing-germanys-open-door))

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, *115*(49), 12441–12446. Retrieved from https://www.pnas.org/content/115/49/12441 doi: 10.1073/pnas.1808083115

Cowan, S. K., & Baldassarri, D. (2018). "It could turn ugly": Selective disclosure of attitudes in political discussion networks. *Social Networks*, *52*, 1–17.

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359.

Crandall, C. S., Ferguson, M. A., & Bahns, A. J. (2013). When we see prejudice: The normative window and social change. In C. S. Crandall & C. Stangor (Eds.), *Frontiers of social psychology. stereotyping and prejudice* (pp. 53–69). New York, NY: Psychology Press.

Crandall, C. S., Miller, J. M., & White, M. H. (2018). Changing norms following the 2016 U.S. presidential election: The trump effect on prejudice. *Social Psychological and Personality Science*, *9*(2), 186–192.

Crandall, C. S., & Stangor, C. (2005). Conformity and prejudice. In J. F. Dovidio, P. Glic, & L. Rudman (Eds.), *On the nature of prejudice: Fifty years after allport* (pp. 295–309). Malden, MA: Blackwell Publishing.

Czymara, C. S., & Schmidt-Catran, A. W. (2017). Refugees unwelcome? changes in the public acceptance of immigrants and refugees in germany in the course of europe's 'immigration crisis'. *European Sociological Review*, *33*(6), 735–751.

DellaVigna, S., List, J. A., Malmendier, U., & Rao, G. (2016). Voting to tell others. *The Review of Economic Studies*, *84*(1), 143–181.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629.

Dimant, E. (2018). Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Available at SSRN 3286770*.

Disha, I., Cavendish, J. C., & King, R. D. (2011). Historical events and spaces of hate: Hate crimes against Arabs and Muslims in post-9/11 America. *Social Problems*, *58*(1), 21–46.

Dovidio, J. F., & Gaertner, S. L. (1986). *Prejudice, discrimination, and racism: Historical trends and contemporary approaches.* Academic Press.

Dovidio, J. F., & Gaertner, S. L. (1991). Changes in the expression and assessment of racial prejudice. In H. J. Knopke, R. J. Norrell, & R. W. Rogers (Eds.), *Opening doors: Perspectives on race relations in contemporary america* (pp. 119–225). Tuscaloosa: The University of Alabama Press.

Duckitt, J. H. (1992). Psychology and prejudice: A historical analysis and integrative framework. *American Psychologist*, *47*(10), 1182.

Durkheim, E. (1897). *Suicide: A study in sociology.* New York, NY: Free Press.

Echebarria-Echabe, A., & Fernández-Guede, E. (2006). Effects of terrorism on attitudes and ideological orientation. *European Journal of Social Psychology*, *36*(2), 259–265.

Elster, J. (1989). Social norms and economic theory. *The Journal of Economic Perspectives*, *3*(4), 99–117.

European Commission against Racism and Intolerance. (2016, March). *Recommendation no. 15 on combating hate speech, adopted on december 2015* (General Policy Recommendation). Strasbourg: Council of Europe.

European Union Agency for Fundamental Rights. (2016, October). *Current migration situation in the eu: hate crime* (Tech. Rep.). FRA. (Retrieved from `https://fra.europa.eu/en/publication/2016/current-migration-situation-eu-hate-crime`)

Europol. (2017). *EU Terrorism situation and trend report (te – sat)* (Vol. 11; Tech. Rep.). The Hague, Netherlands: Europol, EU. (Retrieved from `https://www.europol.europa.eu/activities-services/main-reports/eu-terrorism-situation-and-trend-report-te-sat-2017`)

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American economic review*, *90*(4), 980–994.

Ford, R. (2008). Is racial prejudice declining in Britain? *The British Journal of Sociology*, *59*(4), 609–636.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* UNESCO Publishing.

Geraci, M., & Bottai, M. (2014). Linear quantile mixed models. *Statistics and computing*, *24*(3), 461–479.

Givens, T. E. (2004). The radical right gender gap. *Comparative Political Studies*, *37*(1), 30–54.

Goodman, E., & Cherubini, F. (2013). Online comment moderation: emerging best practices. *Germany: Darmstadt, The World Association of Newspapers WAN-IFRA. Http://www. wan-ifra. org/reports/2013/10/04/online-comment-moderation-emerging-best-practices (17.9. 2014)*.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision*

Making, *26*(3), 213–224.

Greussing, E., & Boomgaarden, H. G. (2017). Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of Ethnic and Migration Studies*, *43*(11), 1749–1774.

Hack, J. (2016, July). *German train ax–attack puts Merkel migrant policy back in spotlight. The Guardian*, July 18. (Retrieved July, 2016 (`https://www.reuters.com/article/us-europe-attacks-germany/german-train-ax-attack-puts-merkel-migrant-policy-back-in-spotlight-idUSKCN0ZY2LA`))

Hanes, E., & Machin, S. (2014). Hate crime in the wake of terror attacks: Evidence from 7/7 and 9/11. *Journal of Contemporary Criminal Justice*, *30*(3), 247–267.

Harteveld, E., & Ivarsflaten, E. (2016). Why women avoid the radical right: Internalized norms and party reputations. *British Journal of Political Science*, 1–16.

Harteveld, E., & Ivarsflaten, E. (2018). Why women avoid the radical right: Internalized norms and party reputations. *British Journal of Political Science*, *48*(2), 369–384.

Harteveld, E., Van Der Brug, W., Dahlberg, S., & Kokkonen, A. (2015). The gender gap in populist radical-right voting: examining the demand side in western and eastern europe. *Patterns of Prejudice*, *49*(1-2), 103–134.

Hechter, M., & Opp, K.-D. (2001). *Social norms.* New York: Russell Sage Foundation.

Heckathorn, D. D. (1988). Collective sanctions and the creation of prisoner's dilemma norms. *American Journal of Sociology*, *94*(3), 535–562.

Hedström, P., Swedberg, R., & Hernes, G. (1998). *Social mechanisms: An analytical approach to social theory.* Cambridge University Press.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual review of sociology*, *36*, 49–67.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond weird: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, *33*(2-3), 111–135.

Henson, B., Reyns, B. W., & Fisher, B. S. (2013). Fear of crime online? examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. *Journal of Contemporary Criminal Justice*, *29*(4), 475–497.

Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, *6*(3), 89–112.

Hirvonen, K. (2013). Sweden: when hate becomes the norm. *Race & Class*, *55*(1), 78–86.

Hoffman, M. A., & Bearman, P. S. (2015). Bringing anomie back in: Exceptional events and excess suicide. *Sociological Science*, *2*, 186–210.

Hogg, M. A., Sherman, D. K., Dierselhuis, J., Maitner, A. T., & Moffitt, G. (2007). Uncertainty, entitativity, and group identification. *Journal of Experimental Social Psychology*, *43*(1), 135–142.

Horne, C. (2009). *The rewards of punishment: A relational theory of norm enforcement.* Stanford University Press.

Horne, C., Tinkler, J., & Przepiorka, W. (2018). Behavioral regularities and norm stickiness: The cases of transracial adoption and online privacy. *Social Research: An International*

Quarterly, *85*(1), 93–113.

Hövermann, A., Messner, S. F., & Zick, A. (2015). Anomie, marketization, and prejudice toward purportedly unprofitable groups: Elaborating a theoretical approach on anomie-driven prejudices. *Acta Sociologica*, *58*(3), 215–231.

Huddy, L., & Feldman, S. (2009). On assessing the political effects of racial prejudice. *Annual Review of Political Science*, *12*, 423–447.

Huddy, L., Khatib, N., & Capelos, T. (2002). Trends: Reactions to the terrorist attacks of September 11, 2001. *The Public Opinion Quarterly*, *66*(3), 418–450.

Ignazi, P. (1992). The silent counter-revolution: Hypotheses on the emergence of extreme right-wing parties in europe. *European Journal of Political Research*, *22*(1), 3–34.

Immerzeel, T., Coffé, H., & Van der Lippe, T. (2015). Explaining the gender gap in radical right voting: A cross-national investigation in 12 western european countries. *Comparative European Politics*, *13*(2), 263–286.

Ivarsflaten, E., Blinder, S., & Ford, R. (2010). The anti-racism norm in western european immigration politics: Why we need to consider it and how to measure it. *Journal of Elections, Public Opinion and Parties*, *20*(4), 421–445.

Jäckle, S., & König, P. D. (2018). Threatening events and anti-refugee violence: An empirical analysis in the wake of the refugee crisis during the years 2015 and 2016 in germany. *European Sociological Review*.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625.

Kalter, F., & Kroneberg, C. (2014). Between mechanism talk and mechanism cult: new emphases in explanatory sociology and empirical research. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *66*(1), 91–115.

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*(5908), 1681–1685.

Kennedy, M. A., & Taylor, M. A. (2010). Online harassment and victimization of college students. *Justice Policy Journal*, *7*(1), 1–21.

Keuschnigg, M., Bader, F., & Bracher, J. (2016). Using crowdsourced online experiments to study context-dependency of behavior. *Social science research*, *59*, 68–82.

Keuschnigg, M., & Wolbring, T. (2015). Disorder, social capital, and norm violation: Three field experiments on the broken windows thesis. *Rationality and Society*, *27*(1), 96–126.

King, R. D., & Sutton, G. M. (2013). High times for hate crimes: Explaining the temporal clustering of hate-motivated offending. *Criminology*, *51*(4), 871–894.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, *9*, 155–176.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, *15*(4), 143–156.

Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., ... Riedl, J. (2012). *Building successful online communities: Evidence-based social design.* Mit Press.

Kreis, R. (2017). # refugeesnotwelcome: Anti-refugee discourse on Twitter. *Discourse & Communication*, *11*(5), 498–514.

Krupka, E., & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, *30*(3), 307–320.

Krupka, E., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.

Kuran, T. (1995). *Private truths, public lies: The social consequences of preference falsification.* Cambridge, Massachusetts: Harvard University Press.

Lecheler, S., & De Vreese, C. H. (2011). Getting real: The duration of framing effects. *Journal of Communication*, *61*(5), 959–983.

Legewie, J. (2013). Terrorist events and attitudes toward immigrants: A natural experiment. *American Journal of Sociology*, *118*(5), 1199–1245.

Little, D. (1991). *Varieties of social explanation* (Vol. 141). Boulder: Westview Press.

Little, D. (2011). Causal mechanisms in the social realm. *Causality in the Sciences*, *27395*.

Mantilla, K. (2013). Gendertrolling: Misogyny adapts to new media. *Feminist Studies*, *39*(2), 563–570.

Manzo, G. (2014). *Analytical sociology: actions and networks.* John Wiley & Sons.

Matias, J. N. (2016). *Posting rules in online discussions prevents problems & increases participation.*

Mendelberg, T. (2001). *The race card: Campaign strategy, implicit messages, and the norm of equality.* Princeton, Oxford: Princeton University Press.

Mize, T. D. (2019). Best practices for estimating, interpreting, and presenting nonlinear interaction effects. *Sociological Science*, *6*, 81–117.

Mudde, C. (2004). The populist zeitgeist. *Government and opposition*, *39*(4), 541–563.

Müller, K., & Schwarz, C. (2017, March). *Fanning the flames of hate: Social media and hate crime.* (Available at SSRN: `https://ssrn.com/abstract=2918883`)

Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 1–21.

Myers, D. G. (1982). Polarizing effects of social interaction. *Group Decision Making*, *14*, 125–161.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and social psychology bulletin*, *34*(7), 913–923.

Norris, P., & Inglehart, R. (2019). *Cultural backlash: Trump, brexit, and authoritarian populism.* Cambridge University Press.

Oltermann, P. (2016, July). *Germany's first attack by radicalised asylum seeker alarms officials. The Guardian*, July 19. (Retrieved June, 2017 (`https://https://www.theguardian.com/`

`world/2016/jul/19/germany-train-attack-could-prompt-rethink-of-counter-terrorism-policy`))

Opp, K.-D. (2004). " what is is always becoming what ought to be." how political action generates a participation norm1. *European Sociological Review*, *20*(1), 13–29.

Paluck, E. L. (2009a). Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *Journal of Personality and Social Psychology*, *96*(3), 574.

Paluck, E. L. (2009b). What's in a norm? sources and processes of norm change. *Journal of Personality and Social Psychology*, *96*, 594–600.

Paluck, E. L. (2011). Peer pressure against prejudice: A high school field experiment examining social network change. *Journal of Experimental Social Psychology*, *47*(2), 350–358.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339–367.

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, *113*(3), 566–571.

Parigi, P., Santana, J. J., & Cook, K. S. (2017). Online field experiments: studying social interactions in context. *Social Psychology Quarterly*, *80*(1), 1–19.

Pettigrew, T. F. (1958). Personality and sociocultural factors in intergroup attitudes: A cross-national comparison. *Journal of Conflict Resolution*, 29–42.

Pettigrew, T. F. (1991). Normative theory in intergroup relations: Explaining both harmony and conflict. *Psychology & Developing Societies*, *3*(1), 3–16.

Ratcliff, J. J., Lassiter, G. D., Markman, K. D., & Snyder, C. J. (2006). Gender differences in attitudes toward gay men and lesbians: The role of motivation to respond without prejudice. *Personality and Social Psychology Bulletin*, *32*(10), 1325–1338.

Rauhut, H., & Winter, F. (2012). On the validity of laboratory research in the political and social sciences: The example of crime and punishment. In B. Kittel, W. J. Luhan, & R. B. Morton (Eds.), *Experimental political science: Principles and practices* (pp. 209–232). London: Palgrave Macmillan UK.

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104.

Riek, B. M., Mania, E. W., & Gaertner, S. L. (2006). Intergroup threat and outgroup attitudes: A meta-analytic review. *Personality and Social Psychology Review*, *10*(4), 336–353.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016, sep). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Eds.), *Proceedings of NLP 4 CMC III: 3rd Workshop on natural language processing for computer-mediated communication* (Vol. 17, pp. 6–9). Bochum.

Rost, K., Stahel, L., & Frey, B. S. (2016). Digital social norm enforcement: Online firestorms in social media. *PLoS one*, *11*(6).

Rydgren, J. (2005). Is extreme right-wing populism contagious? explaining the emergence of a new party family. *European Journal of Political Research*, *44*(3), 413–437.

Rydgren, J. (2018). *The oxford handbook of the radical right.* Oxford University Press.

Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan* (pp. 1–23).

Schuman, H., Steeh, C., Bobo, L., & Krysan, M. (1997). *Racial attitudes in America: Trends and interpretations.* Cambridge, MA: Harvard University Press.

Sechrist, G. B., & Stangor, C. (2005). Prejudice as social norms. In *Social psychology of prejudice: historical and contemporary issues/edited by christian s. crandall, mark schaller. lawrence, kan.: Lewinian press, c2005.* Lawrence, Kan.: Lewinian Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Wadsworth Cengage learning.

Shapiro, J. R., & Neuberg, S. L. (2008). When do the stigmatized stigmatize? the ironic effects of being accountable to (perceived) majority group prejudice-expression norms. *Journal of personality and social psychology*, *95*(4), 877.

Sherif, M., & Sherif, C. W. (1953). *Groups in harmony and tension; an integration of studies of intergroup relations.* New York: Harper & Brothers.

Sliwka, D. (2007). Trust as a signal of a social norm and the hidden costs of incentive schemes. *American Economic Review*, *97*(3), 999–1012.

Stangor, C., Sechrist, G. B., & Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, *27*(4), 486–496.

Stürmer, S., Rohmann, A., Froehlich, L., & van der Noll, J. (2018). Muslim immigration, critical events, and the seeds of majority members' support for radical responses: An interactionist perspective. *Personality and Social Psychology Bulletin*.

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *10*(1), 181–211.

Traugott, M., Brader, T., Coral, D., Curtin, R., Featherman, D., Groves, R., . . . others (2002). How Americans responded: A study of public reactions to 9/11/01. *Political Science & Politics*, *35*(3), 511–516.

Ullmann-Margalit, E. (2015). *The emergence of norms.* OUP Oxford.

Voss, T. (2001). *Game-theoretical perspectives on the emergence of social norms.* na.

Walker, B. H., Sinclair, H. C., & MacArthur, J. (2015). Social norms versus social motives: The effects of social influence and motivation to control prejudiced reactions on the expression of prejudice. *Social Influence*, *10*(1), 55–67.

Walters, M., Brown, R., & Wiedlitzka, S. (2016, July). *Causes and motivations of hate crime* (Research Report No. 102). (Available at SSRN: `https://ssrn.com/abstract=2918883`)

Weber, C. R., Lavine, H., Huddy, L., & Federico, C. M. (2014). Placing racial stereotypes in context: Social desirability and the politics of racial hostility. *American Journal of Political Science*, *58*(1), 63–78.

Wike, R., Stokes, B., & Simmons, K. (2016). Europeans fear wave of refugees will mean more terrorism, fewer jobs. *Pew Research Center*, *11*, 2016.

Willer, R., Kuwabara, K., & Macy, M. W. (2009). The false enforcement of unpopular norms. *American Journal of Sociology*, *115*(2), 451–490.

Williams, M. L., & Burnap, P. (2015). Cyberhate on social media in the aftermath of Woolwich: A aase study in computational criminology and big data. *British Journal of Criminology*, *56*(2), 211–238.

Winter, F., Rauhut, H., & Miller, L. (2017). Dynamic bargaining and normative conflict. *Max-Planck-Institute for Research on Collective Goods Working Paper.*.

Xiao, E., & Bicchieri, C. (2010). When equality trumps reciprocity. *Journal of Economic Psychology*, *31*(3), 456–470.

Zitek, E. M., & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, *43*(6), 867–876.

Author's Biography

Amalia Álvarez Benjumea is a research fellow at the Max Planck Institute for Research on Collective Goods in Bonn, Germany, where she works under the supervision of Fabian Winter in the Max Planck research group"Mechanisms for Normative Change". Since 2016, she has also been a graduate student in sociology at the Faculty of Management, Economics and Social Sciences at the University of Cologne. Before coming to Germany, she gained a Bachelor's degree in sociology from the Autonomous University of Barcelona, in Spain. She also completed a Master of Science in Social Cognition at University College London (UCL), and a Master of Science in Applied Social Research at Stockholm University (Sweden). Furthermore, she was a visiting fellow in the Norms and Networks Cluster (NNC) at the University of Groningen, The Netherlands, and at the Department of Sociology at New York University (NYU). Her research interests include methods of online experiments, the effect of social information on normative behavior, the perception of norms, and opinion dynamics under normative and social influence.

Amalia Álvarez Benjumea

Kurt-Schumacher-Str. 10
D-53113 Bonn
Germany

Phone: (0049) 228 91416-856
Mobile: (0049) 176 600 26 652
alvarezbenjumea@coll.mpg.de

Current Position

Research Fellow, Max Planck Institute for Research on Collective Goods, Bonn (Germany).

MPRG*"Mechanisms of Normative Change"*
P.I.: Fabian Winter

Visiting Positions

Visiting researcher, Department of Sociology, New York University. April-July 2018. Host: Delia Baldassarri

Visiting researcher, Norms and Networks Cluster (NNC) at University of Groningen (RUG). February 2018. Host: Andreas Flache

Education

PhD Candidate in Sociology, Cologne Graduate School in Management, Economics and Social Sciences, University of Cologne, *Expected 2019.*
Dissertation project: "The spreading of hostility: unraveling of social norms of communication"

Supervisor: Clemens Kroneberg

International Max Planck Research Graduate School Adapting behaviour in a fundamentally uncertain world, 2018

M.Sc. Applied Social Research, Department of Sociology, Stockholm University, 2015.
Master's thesis: *"Friendship choices and ethnic homophily"*
Supervisor: Jens Rydgren

M.Sc. Social Cognition: research and motivations, Division of Psychology & Language Sciences, University College London, 2011.

B.Sc. Sociology, Faculty of Sociology and Political Science, Autonomous University of Barcelona, 2010.

Research Interests

My research interests include the effect of social information on normative behaviour, perception of norms, opinion dynamics under normative and social influence, and experimental methods, especially online experiments.

Publications in Peer-Reviewed Journals

Álvarez-Benjumea, A., & Winter, F. (2018). Normative Change and Culture of Hate: An Experiment in Online Environments. *European Sociological Review*, 34(3), 223-237.

Working Papers

The Breakdown of Anti-Racist Norms: A Natural Experiment on Normative Uncertainty after Terrorist Attacks (together with Winter, F.)

Public Signals as Coordination Devices: the Moderating Effect of Group Identity (together with Freund, L., Luckner, K., and Winter, F.)

Work in Progress

Uncovering Hidden Opinions: The Effect of Social Acceptability on Disclosure of Anti-immigrant views

Gender Differences in Reaction to an Anti-egalitarian Descriptive Norm: When Normative Cues Backfire.

Selected Conferences / Workshops

Oral Presentations in Conferences

3. *Normative change and culture of hate: a randomized experiment in online communities*, 10th Conference of the International Network of Analytical Sociologists (INAS), 8th-9th June, 2017, Oslo (Norway)

2. *Normative change and culture of hate: a randomized experiment in online communities*, 10th JDM Meeting, Max Planck Institute for Research on Collective Goods, 1st-2nd June, Bonn (Germany)

1. *Friendship Choices and ethnic background*, XII Spanish Sociology Congress. Spanish Sociological Federation (FES), 30th July, 2016, Gijón (Spain)

Oral Presentations in Workshops

4. I Jornadas *Experimentos en Sociología y política* at Universidad Pablo de Olavide, 28th-29st January, 2019, Sevilla, Spain.

3. Workshop on *Rationality and Irrationality in Democracy* at the XIX ISA World Congress of Sociology, 15th-21st July, 2018, Toronto, Canada.

2. PhD workshop at the 13th conference of the European sociological association (ESA) , 26th-28th August, 2017, Athens, Greece.

1. Workshop on *Common good and self-interest in the digital Society* at the congress of the Swiss sociological association (SSA), 23th-24th June, 2017, Zurich, Switzerland.

Invited talks/Seminars

1. *When do terrorist attacks increase hate? Evidence from a Natural Experiment.* Democracy, Elections and Citizenship (DEC) Seminar at the Department of Political Science of the Autonomous University of Barcelona, 29th November, 2018, Spain.

Organization of Scientific Meetings

1. Workshop on *Mechanisms of normative change* at the XIX ISA World Congress of Sociology,15th-21st July, 2018, Toronto, Canada.

Selected courses and Summer Schools

5. *OII Summer Doctoral Programme (SDP 2017)* from the Oxford Internet Institute, University of Oxford, 3rd July - 14th July 2017

4. *10th IMPRS Uncertainty Summer school* from the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World (Uncertainty School), Jena, 24th July-19th August 2016.

3. *Topics in Management and Applied Microeconomics (Empirical Political Economy)*, University of Bonn, 2015/16.

2. *9th IMPRS Uncertainty Summer school* from the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World (Uncertainty School), Jena, 19th July-21st August 2015.

1. *Systematic Reviews and meta analysis*, Stockholm University, January 2015

Past academic positions

Research assistant *"Seminari d'estudis de la Dona"* Autonomous University of Barcelona, Bellaterra (Barcelona), October 2009 - July 2010

Projects *Time organization in companies of Catalonia, Organic Act for Effective Equality Between Women and Men of 2007.*

Scholarships and Prizes

Leonardo DaVinci Scholarship, 2012.
European Sociological Association (ESA) PhD Summer School scholarship, 2017.

Languages

English: Professional knowledge (European C2 level)

German: Basic knowledge (European A2 level)

Swedish: Fluent knowledge (European B2 level)

Spanish: Mother tongue

Additional information

Reviewer/Referee

Journal of Economic Psychology, American Sociological Review, American Political Science Review

Programming Languages and Scientific Applications

Python, Stata, R, UCINET, NetLogo, Pajek,LaTEX, oTree.

Other Activities

PhD internal representative, Max Planck Society 2016-2017

Volunteer job with EU-migration in *Crossroads*, Stockholm 2013-15
Data analyst

Last updated: May 28, 2020

“Hiermit versichere ich an Eides Statt, dass ich die vorgelegte Disseration selbstaändig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Aussagen, Daten und Konzepte sind unter Angabe del Quelle gekennzeichnet. Beu del Auswahl und Auswertung folgenden Materials Haben mir die nachstehend aufgefürten Personen in der jeweils beschriebenen Weise entgeltlich/unentgeltlich (zutreffendes bitte unterstreichen) geholfen:

Weitere Personen, neben den in der Einleitung der Dissertation aufgeführten Koautorinnen und Koautoren, waren an der inhaltlish-materiellen Erstellung del vorliegenden Dissertation nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Dissertation wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelect. Ich versichere, dass ich nacch bestem Wissem die reine Wahrheit gesagt und nichts verschwiegen habe.”

Amalia Álvarez Benjumea