Normative change and culture of hate: a randomized experiment in online communities

Amalia Álvarez    Fabian Winter

Max Planck Institute for Research on Collective Goods

Venice Rational Choice Sociology 2017
November 22, 2017
Facebook, Twitter, and Google crack down on hate speech in Germany

Facebook Launches New Initiative Against Online Extremism And Hate Speech
Individuals conform to salient norms and cater to the audience (Bicchieri, 2005; Cialdini and Goldstein, 2004).

Individual’s perception of social acceptability of hate speech affects willingness to express it publicly (Bursztyn et al., 2017).

The perception of a social norm depends on the presence of the relevant normative expectations:

- Person believes that a sufficiently large subset of people follows the norm (descriptive norm), or

- Person believes that a sufficiently large subset of people expects her to follow the norm (injunctive norm)
Norms of communication behavior II

Both types of normative expectations have been successfully used to reduce hate/prejudiced speech:

*Descriptive norm:*

- Manipulating consensus information over negative stereotypes reduced the adherence to negative stereotypes (Stangor et al., 2001)
Both types of normative expectations have been successfully used to reduce hate/prejudiced speech:

**Descriptive norm:**
- Manipulating consensus information over negative stereotypes reduced the adherence to negative stereotypes (Stangor et al., 2001)

**Injunctive norm:**
- Individuals are more likely to oppose discrimination immediately after hearing someone else do so first (Cialdini and Trost, 1998; Blanchard et al., 1994)
- Informal peer-sanctions can have a deterrent effect and prevent online hate speech (Munger, 2016)
Updating norm perception in the online forum

**Censoring**

*(Removing hateful content)*

Bias the perceived pattern of behaviour

↓

Others conform to the norm

**H1: (Descriptive Norms Effect)**

Removing examples of hate speech in the online context, will accentuate a descriptive norm and lead to less hostile content.

**Counter-comments** *(Informal sanctions from other users)*

Observing peer punishment signals public/group disapproval

↓

Others expect me to conform to the norm

**H2: (Injunctive Norms Effect)**

Observing verbal sanctions to previous examples of hate speech strongly signals existence of injunctive norm and leads to less hostile content.
Updating norm perception in the online forum

**Censoring**  
*(Removing hateful content)*

Bias the perceived pattern of behaviour

↓

Others conform to the norm

**H1: (Descriptive Norms Effect)**

Removing examples of hate speech in the online context, will accentuate a descriptive norm and lead to less hostile content.

---

**Counter-comments**  
*(Informal sanctions from other users)*

Observing peer punishment signals public/group disapproval

↓

Others expect me to conform to the norm

**H2: (Injunctive Norms Effect)**

Observing verbal sanctions to previous examples of hate speech strongly signals existence of injunctive norm and leads to less hostile content.
Approach 1: Harvesting
1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

Pro: data collection, sentiment analysis, external validity

Con: selection, endogeneity

Approach 2: Experiment
1. Construct your own Facebook
2. Collect comments
3. Construct treatments
4. Invite participants
5. Random assignment to treatments

Pro: identification

Con: internal validity
Approach 1: Harvesting
1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

Pro: data collection, sentiment analysis, external validity
Con: selection, endogeneity

Approach 2: Experiment
1. Construct your own facebook
2. Collect comments
3. Construct treatments
4. Invite participants
5. Random assignment to treatments

Pro: Identification
Con: int. validity
**Approach 1: Harvesting**

1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

**Pro:**
- Data collection,
- Sentiment analysis,
- External validity

**Con:**
- Selection,
- Endogeneity

**Approach 2: Experiment**

1. Construct your own Facebook
2. Collect comments
3. Construct treatments
4. Invite participants
5. Random assignment to treatments

**Pro:**
- Identification

**Con:**
- Internal validity
Approach 1: Harvesting

1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

**Pro:** data collection, sentiment analysis, external validity

**Con:** selection, endogeneity

Approach 2: Experiment

1. Construct your own Facebook
2. Collect comments
3. Construct treatments
4. Invite participants
5. Random assignment to treatments

**Pro:** Identification

**Con:** internal validity
**Approach 1: Harvesting**

1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

**Pro:** data collection, sentiment analysis, external validity

**Con:** selection, endogeneity

**Approach 2: Experiment**

1. Construct your own facebook
2. Collect comments
3. Construct treatments
4. Invite participants
5. **Random assignment to treatments**
Approach 1: Harvesting
1. Find thread with strict censoring rules
2. Find thread with lenient censoring rules
3. Harvest (a ton of) data
4. Use e.g. sentiment analysis
5. Estimate treatment effect

**Pro:** data collection, sentiment analysis, external validity

**Con:** selection, endogeneity

Approach 2: Experiment
1. Construct your own facebook
2. Collect comments
3. Construct treatments
4. Invite participants
5. **Random assignment to treatments**

**Pro:** Identification

**Con:** int. validity
Social Media in the Sandbox
Preliminary pictures (N=200)

Online survey: selection of pictures” (N=9)

Online comments collection

Comments classification

Countercomments

Treatments construction

ONLINE EXPERIMENT

Comments classification
Creating an online discussion forum on social topics

- 9 different pictures from 4 different topics
- Users remain anonymous and are given an avatar and an username to use them in the discussion
- 6 comments are displayed: a mix of neutral, positive/friendly and negative/hostile comments (Example comments)
Censoring treatment

- The negative/hostile comments are deleted

- Two treatments:
  - censored
  - extremely censored
Counterspeaking treatment

- Same pictures and same pool of comments
- Hostile comments are now countercommented
Rating of comments

Hate speech score:
Is the comment friendly or hostile towards the group represented in the picture? (Indicate from 1 to 9 where 1 means very friendly and 9 means very hostile)

Hate Speech Indicator:
▶ negative stereotypes
▶ racist slurs
▶ demeaning language
▶ encourages violence
▶ sexist slurs
▶ stigmatizes gender or sexual orientation

Ist der Kommentar freundlich oder feindlich gegenüber der im Foto dargestellten Gruppe?

Welche Merkmale treffen auf den Kommentar zu?

Beinhaltet negative Vorurteile
Beinhaltet 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

Weiter
Rating of comments

**Hate speech score:**

Is the comment friendly or hostile towards the group represented in the picture? *(Indicate from 1 to 9 where 1 means very friendly and 9 means very hostile)*
Rating of comments

**Hate speech score:**
Is the comment friendly or hostile towards the group represented in the picture? *(Indicate from 1 to 9 where 1 means very friendly and 9 means very hostile)*

**Hate Speech Indicator:**
Which of the characteristics applies to the comment?

- negative stereotypes
- racist slurs
- demeaning language
- encourages violence
- sexist slurs
- stigmatizes gender or sexual orientation
Results
Hate speech score by topic and treatment

Figure: Average hate speech score across treatments and topics
Figure: Density distribution of average hate speech score across treatments. The graph depicts the 1 to 9 score scale.

Participants are slightly more prone to use strongly hateful language in the extremely censored treatment.
Hate Speech Indicator

Figure: Proportion of comments that were labeled as hate speech across treatments. Error bars at 95% CI.
Contributions

This project represents a step forward in the research on online hate speech. The results can help design better informed interventions to tackle online hate speech using social norms perception as means for normative change.

Our results add to the literature of social norms, it presents empirical evidence of the effect of social norm perception on the willingness to engage in online hate speech, even in anonymous contexts, without direct punishment, and controlling for selection effects.
Spillover Effects of Hate Speech: A Natural Experiment
Motivation

Do external events such as terrorist attacks lead to changed social norms regarding the expression of hateful views?
External shocks effects on hate expression

A spread of hostile and hate expression is normally linked to terrorist attacks:

- Public expression of anti-foreign sentiment following attacks (Legewie, 2013; Hanes and Machin, 2014).
- Increase in hate speech in online social media context (Williams and Burnap, 2015).

**Breakdown of modesty norms towards may spill over to modesty norms towards other minority groups.** Spillovers: Increases in public expression of hate against some minorities might also spur increases in expression of hate against other minorities (CSBS, 2017).

Change in norms?
External shocks effects on hate expression: mechanisms

Individuals conform to salient norms (Bicchieri, 2005; Cialdini and Goldstein, 2004) and avoid expressing unpopular opinions (Bursztyn et al., 2016).

Individual’s perception of social acceptability of hate speech affects willingness to express it publicly (Bursztyn et al., 2017).

External shocks (e.g. reactions to terrorist attacks...) might induce changes in the social acceptability of certain (extreme) opinions and in the likelihood that these opinions are publicly expressed.
Attack associated with refugees

Refugees/Asylum Seekers

Feminism/Transgender/Gay Rights
Data selection

- We selected data from the baseline and the censored treatments.
- We selected data from the refugees, feminism and LGBT comments threads. The feminist and LGBT comments were merged in a new "Other" category.
- The comments were rated by 577 different raters. Raters were asked to rate 30 comment.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Wave 1</th>
<th></th>
<th>Wave 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Refugees</td>
<td>Other</td>
<td>Refugees</td>
<td>Other</td>
</tr>
<tr>
<td>Baseline</td>
<td>135</td>
<td>227</td>
<td>135</td>
<td>228</td>
</tr>
<tr>
<td>Censored</td>
<td>136</td>
<td>225</td>
<td>135</td>
<td>226</td>
</tr>
<tr>
<td>Extremely censored</td>
<td>134</td>
<td>225</td>
<td>123</td>
<td>204</td>
</tr>
<tr>
<td>Total</td>
<td>405</td>
<td>677</td>
<td>393</td>
<td>658</td>
</tr>
</tbody>
</table>
Results
Change in mean hate speech score: baseline treatment

Figure: Error bars at 95%. The graph displays the changes in hate speech score before and after the terrorist attacks for the two categories. Error bars at 95%.
Change in mean hate speech score

Figure: Error bars at 95%. Left: changes in hate speech score before and after the terrorist attacks for the other category. Right: refugees.
Increase of hate speech towards refugees
Normative change and culture of hate: Summary

- Hate speech is context dependent: **Descriptive norms matter.**
- Censoring extremely negative comments leads to **less hate speech.**
- Extreme Censoring may lead to **more extreme comments.**
- External shocks may erode hate speech norms.
- Weak spillover of hate speech to **unrelated domains**
- Interventions in one domain may inhibit negative dynamics in other domains
Appendix
Treatment Effects - Study 1

\[ Y_{ij} = \beta_0 + \beta_1 \text{Treatment} + \beta_2 \text{Topic} + u_j + \epsilon_{ij} \]

<table>
<thead>
<tr>
<th>Hate speech score:</th>
<th>Model (1)</th>
<th>Model (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>4.61 (0.11)**</td>
<td>4.41 (0.13)**</td>
</tr>
<tr>
<td>Counter-speaking</td>
<td>-0.13 (0.16)</td>
<td>-0.14 (0.15)</td>
</tr>
<tr>
<td>Censored</td>
<td>-0.38 (0.16)*</td>
<td>-0.39 (0.15)*</td>
</tr>
<tr>
<td>Extremely censored</td>
<td>-0.38 (0.16)*</td>
<td>-0.40 (0.16)*</td>
</tr>
<tr>
<td>LGBT</td>
<td>-0.00 (0.09)</td>
<td></td>
</tr>
<tr>
<td>Refugees</td>
<td>0.65 (0.09)**</td>
<td></td>
</tr>
<tr>
<td>Feminism</td>
<td>0.03 (0.09)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1469</td>
<td>1469</td>
</tr>
<tr>
<td>Number of Subjects</td>
<td>180</td>
<td>180</td>
</tr>
</tbody>
</table>

Significance levels: *** \( p < 0.001 \), ** \( p < 0.01 \), * \( p < 0.05 \), † \( p < 0.1 \)

Table: Results from multilevel random models of hate speech score. Model 1 shows main effects of treatments. Model 2 shows main effects of treatments and topics. The baseline treatment and the topic poverty serve as the reference categories.
Increase in hate speech score in time 2 by topic

<table>
<thead>
<tr>
<th></th>
<th>Other</th>
<th>Refugees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.11 (0.16)**</td>
<td>3.90 (0.17)**</td>
</tr>
<tr>
<td>Time 2</td>
<td>0.24 (0.23)</td>
<td>0.56 (0.25)*</td>
</tr>
<tr>
<td>Obs.</td>
<td>455</td>
<td>270</td>
</tr>
<tr>
<td>Groups:Subjects</td>
<td>94</td>
<td>90</td>
</tr>
<tr>
<td>Var: Subjects</td>
<td>0.91</td>
<td>0.86</td>
</tr>
</tbody>
</table>

*** p < 0, ** p < 0.01, * p < 0.05, . p < 0.1

Table: Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Hate speech score
### Increase in hate speech score by topic

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.11 (0.15)**</td>
</tr>
<tr>
<td>Time 2</td>
<td>0.24 (0.22)</td>
</tr>
<tr>
<td>Refugees</td>
<td>0.78 (0.14)**</td>
</tr>
<tr>
<td>Time*Refugees</td>
<td>0.36 (0.20)</td>
</tr>
<tr>
<td>Obs.</td>
<td>725</td>
</tr>
<tr>
<td>Groups:Subjects</td>
<td>94</td>
</tr>
<tr>
<td>Var:Subjects</td>
<td>0.75</td>
</tr>
</tbody>
</table>

* *** $\rho < 0.001$, ** $\rho < 0.01$, * $\rho < 0.05$, . $\rho < 0.1$*

Table: Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Hate speech score
Interaction with experimental treatments

<table>
<thead>
<tr>
<th>Main effects</th>
<th>Model 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.11 (0.15)**</td>
</tr>
<tr>
<td>Refugees</td>
<td>0.78 (0.14)**</td>
</tr>
<tr>
<td>Time 2</td>
<td>0.24 (0.21)</td>
</tr>
<tr>
<td>Censored</td>
<td>−0.14 (0.21)</td>
</tr>
<tr>
<td>Extremely censored</td>
<td>−0.24 (0.21)</td>
</tr>
<tr>
<td>Interaction effects</td>
<td></td>
</tr>
<tr>
<td>Refugees*Time 2</td>
<td>0.36 (0.20)</td>
</tr>
<tr>
<td>Refugees*Censored</td>
<td>0.22 (0.20)</td>
</tr>
<tr>
<td>Refugees*Extremely Censored</td>
<td>0.32 (0.20)</td>
</tr>
<tr>
<td>Time 2*Censored</td>
<td>−0.11 (0.30)</td>
</tr>
<tr>
<td>Time 2*Extremely Censored</td>
<td>−0.05 (0.31)</td>
</tr>
<tr>
<td>Refugees<em>Time 2</em> Censored</td>
<td>−0.48 (0.28)</td>
</tr>
<tr>
<td>Refugees<em>Time 2</em> Ext. Censored</td>
<td>−0.66 (0.28)*</td>
</tr>
</tbody>
</table>

| Obs.                                  | 2133                         |
| Groups:Subjects                       | 274                          |
| Var: Subjects                         | 0.71                         |

Table: Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models of the Hate speech score by treatment, time and category
Comments example

**Friendly:**

▶ „Ich wünsche mir, für jeden der hier leben und arbeiten will, dass er alle erdenkliche Hilfe bekommt um es für sich und seine Familie realisieren zu können“ (Refugees)

▶ „Super Daumen hoch für diese Leute die den Mumm haben sich der Ignoranz zu stellen” (Feminism)

**Neutral:**

▶ „Generell bin ich dagegen sich in der Öffentlichkeit wild zu küssen. Aber gegen einen Kuss habe ich nichts” (LGBT rights)

▶ „Letzten Endes lediglich ein gewöhnungsbedürftiger Anblick. Bis auf die Gesichtsverschleierung kommt es einer Nonne fast schon gleich” (Refugees)

**Hostile:**

▶ „Einfach nur absurd, dass unsere Politiker all diese Leute einfach ohne Papieren einreisen lassen. Die meisten sind eine Gefahr für uns und unsere Kinder. Und können dann nichtmal abgeschoben oder bestraft werden ... ” (Refugees)

▶ „Meine Toleranz hat Grenzen. Transsexuelle haben beim Militär nichts zu suchen” (Feminism)
Kernel density estimates

Figure: Kernel density estimates for the hate speech score in censored (green line) and extremely censored (blue line) treatments
Non-parametric results

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Mean(Score)</th>
<th>Median(Score)</th>
<th>p vs baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>4.61 (1.30)</td>
<td>4.33</td>
<td></td>
</tr>
<tr>
<td>Counter-speaking</td>
<td>4.48 (1.26)</td>
<td>4.33</td>
<td>0.5483</td>
</tr>
<tr>
<td>Censored</td>
<td>4.24 (1.06)</td>
<td>4</td>
<td>0.0023</td>
</tr>
<tr>
<td>Extremely censored</td>
<td>4.24 (1.24)</td>
<td>4</td>
<td>0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>4.39</td>
<td>4.33</td>
<td></td>
</tr>
</tbody>
</table>

Table: Mean and median hate speech score in the different treatments (standard deviation in parentheses). In column 4, the level of hate speech in the treatments is compared to the baseline level (Kruskal-Wallis test).
Models

\[ Y_{ij} = \beta_0 + \beta_1 \text{Treatment} + u_j + \epsilon_{ij} \quad (1) \]

\[ Y_{ij} = \beta_0 + \beta_1 \text{Treatment} + \beta_2 \text{Topic} + u_j + \epsilon_{ij} \quad (2) \]

\[ Y_{ij} = \beta_0 + \beta_1 \text{Topic} + \beta_2 (\text{Treatment} \ast \text{Topic}) + u_j + \epsilon_{ij} \quad (3) \]
# Results from multilevel random models of hate speech score

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>4.61 (0.11)**</td>
<td>4.41 (0.13)**</td>
<td>4.20 (0.18)**</td>
</tr>
<tr>
<td>Counter-speaking</td>
<td>-0.13 (0.16)</td>
<td>-0.14 (0.15)</td>
<td></td>
</tr>
<tr>
<td>Censored</td>
<td>-0.38 (0.16)*</td>
<td>-0.39 (0.15)*</td>
<td></td>
</tr>
<tr>
<td>Extremely censored</td>
<td>-0.38 (0.16)*</td>
<td>-0.40 (0.16)*</td>
<td></td>
</tr>
<tr>
<td>LGBT</td>
<td>-0.00 (0.09)</td>
<td></td>
<td>0.21 (0.18)</td>
</tr>
<tr>
<td>Refugees/Multiculturality</td>
<td>0.65 (0.09)**</td>
<td>1.01 (0.18)**</td>
<td></td>
</tr>
<tr>
<td>Feminism</td>
<td>0.03 (0.09)</td>
<td></td>
<td>0.19 (0.17)</td>
</tr>
<tr>
<td><strong>Interaction effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poverty*Counter-speaking</td>
<td></td>
<td></td>
<td>0.08 (0.25)</td>
</tr>
<tr>
<td>Poverty*Censored</td>
<td></td>
<td></td>
<td>-0.02 (0.25)</td>
</tr>
<tr>
<td>Poverty*Extremely</td>
<td></td>
<td></td>
<td>-0.12 (0.26)</td>
</tr>
<tr>
<td>LGBT*Counter-speaking</td>
<td></td>
<td></td>
<td>-0.16 (0.20)</td>
</tr>
<tr>
<td>LGBT*Censored</td>
<td></td>
<td></td>
<td>-0.33 (0.20)</td>
</tr>
<tr>
<td>LGBT*Extremely</td>
<td></td>
<td></td>
<td>-0.44 (0.21)*</td>
</tr>
<tr>
<td>Refugees*Counter-speaking</td>
<td></td>
<td></td>
<td>-0.26 (0.19)</td>
</tr>
<tr>
<td>Refugees*Censored</td>
<td></td>
<td></td>
<td>-0.65 (0.19)**</td>
</tr>
<tr>
<td>Refugees*Extremely</td>
<td></td>
<td></td>
<td>-0.61 (0.19)**</td>
</tr>
<tr>
<td>Feminism*Counter-speaking</td>
<td></td>
<td></td>
<td>-0.11 (0.18)</td>
</tr>
<tr>
<td>Feminism*Censored</td>
<td></td>
<td></td>
<td>-0.34 (0.18)^†</td>
</tr>
<tr>
<td>Feminism*Extremely</td>
<td></td>
<td></td>
<td>-0.28 (0.19)</td>
</tr>
<tr>
<td><strong>Random Parts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groups: Subjects</td>
<td>180</td>
<td>180</td>
<td>180</td>
</tr>
<tr>
<td>Var: Subjects</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>Residual Variance</td>
<td>1.06</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Obs.</td>
<td>1469</td>
<td>1469</td>
<td>1469</td>
</tr>
</tbody>
</table>
Hate Speech Indicator by participant

Figure: Error bars at 95%. Left: Number of participants that made at least one hate comment.
References I


