

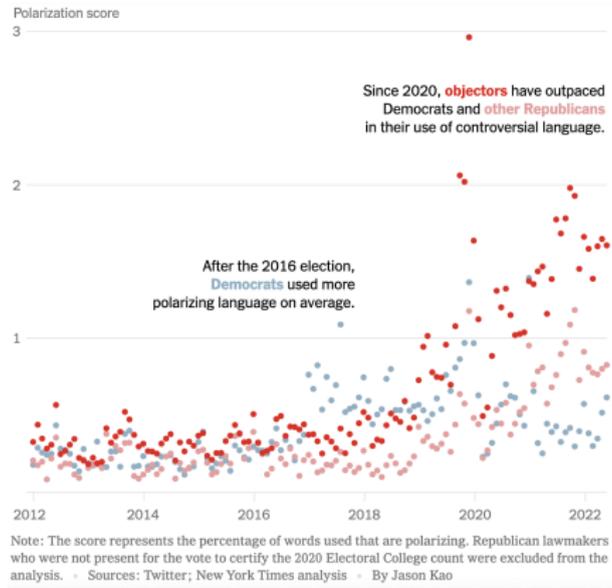
Do Sanctioning and Monitoring Affect Political Elites' Online Toxicity?

Evidence from a Field Experiment
on US General Election Candidates

Yunus Emre Orhan (NDSU)
Val Mechkova (U. Gothenburg)
Dan Pemstein (NDSU)
Brigitte Seim (UNC)
Steven Wilson (Brandeis)

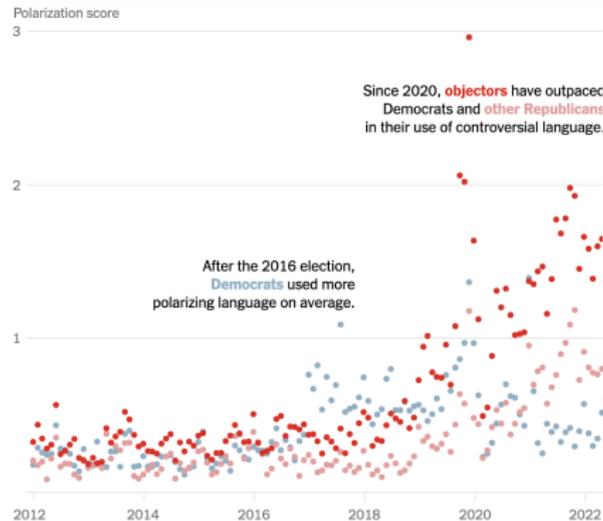
Motivation

- Increasing online polarization and toxicity



Motivation

- Increasing online polarization and toxicity



Madison Cawthorn
@CawthornforNC

.@AOC's top congressional priorities:

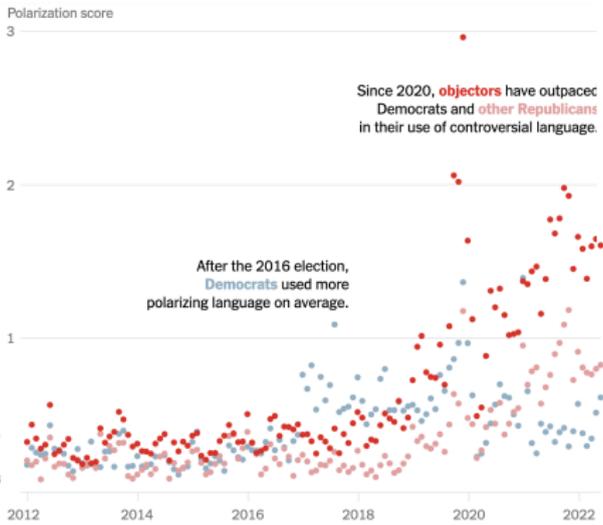
- 1) Photoshoots ✓
- 2) Virtue Signaling ✓
- 3) Destroying America ✓

10:27 AM · Apr 5, 2021 · Twitter for iPhone

Note: The score represents the percentage of words used that are polarizing. Republican lawmakers who were not present for the vote to certify the 2020 Electoral College count were excluded from the analysis. Sources: Twitter; New York Times analysis By Jason Kao

Motivation

- Increasing online polarization and toxicity



Madison Cawthorn ✓
@CawthornforNC

.@AOC's top congressional priorities:

- 1) Photoshoots ✓
- 2) Virtue Signaling ✓
- 3) Destroying America ✓

10:27 AM · Apr 5, 2021 · Twitter for iPhone



Bill Pascrell, Jr. ✓
@BillPascrell

If you're wondering why so many republican candidates for office are blithering idiots know they are an expression of the republican party's contempt for you and American democracy itself.

8:39 PM · Oct 16, 2022 · Twitter for iPhone

Note: The score represents the percentage of words used that are polarizing. Republican lawmakers who were not present for the vote to certify the 2020 Electoral College count were excluded from the analysis. Sources: Twitter; New York Times analysis By Jason Kao

Motivation

- Increasing online polarization and toxicity
- Potential pernicious effects of elites' toxic language
 - Swamping out constructive debate (Druckman, Peterson & Slothuus 2013)
 - Exacerbating polarization (Bail et al. 2018)
 - Encouraging harassment of historically under-represented groups (Mechkova & Wilson 2021)
 - Stimulating political violence (Feuer, Schmidt & Broadwater 2022)

Motivation

- Increasing online polarization and toxicity
- Potential pernicious effects of elites' toxic language
 - Swamping out constructive debate (Druckman, Peterson & Slothuus 2013)
 - Exacerbating polarization (Bail et al. 2018)
 - Encouraging harassment of historically under-represented groups (Mechkova & Wilson 2021)
 - Stimulating political violence (Feuer, Schmidt & Broadwater 2022)
- Can we *cheaply* reduce elites' online toxicity?

Research Questions

- 1 Does bottom-up **social sanctioning** reduce politicians' online toxic speech?
(Rasinski and Czopp, 2010; Munger 2017)
- 2 Does top-down **monitoring** reduce politicians' online toxic speech?
(Grossman and Hanlon, 2014; Grossman and Michelitch, 2018; Nyhan and Reifler, 2015)
- 3 Does top-down **monitoring** reduce politicians' willingness to communicate online (chill speech)?

Contributions

- Adaptation of bottom-up sanctioning to **elite** targets
- Application of monitoring intervention to **online behavior**
- **Comparative analysis** of political elite social media behavior
including Tunisia, Turkey, Brazil, India, Italy, the Philippines,
France, and Australia.

Data

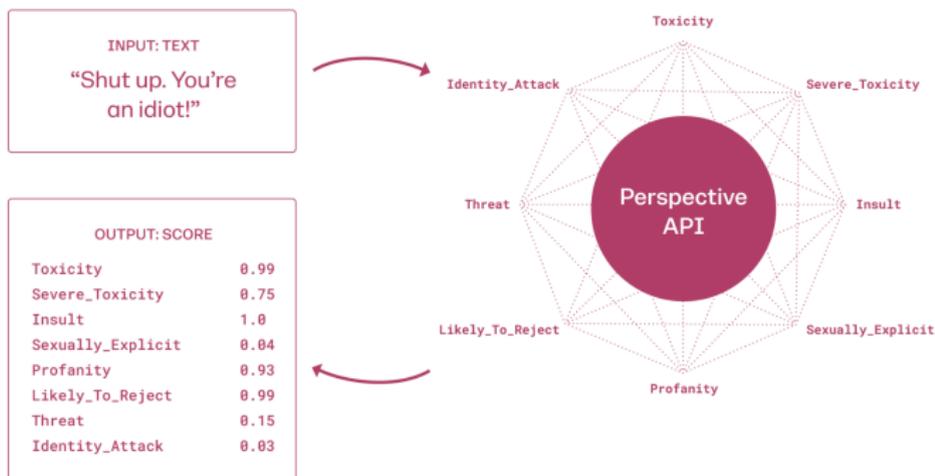
- Country: The US
- Period: Oct-13-2022 - Dec-15-2022
- Data: 2022 Midterm Election Candidates
- Sample size: 3560
- Inclusion Criteria:
 - 1 General election candidate for legislative office (Fed/State)
 - 2 Affiliation with Democratic or Republican Party
 - 3 Twitter account

Hypotheses

- H1: Top-down **monitoring** will decrease toxicity
- H2: Bottom-up **social sanctioning** will decrease toxicity
- H3: Top-down **monitoring** will decrease toxicity more than bottom-up **social sanctioning**
- H4: Top-down **monitoring** will decrease tweet frequency (number of tweets per week)

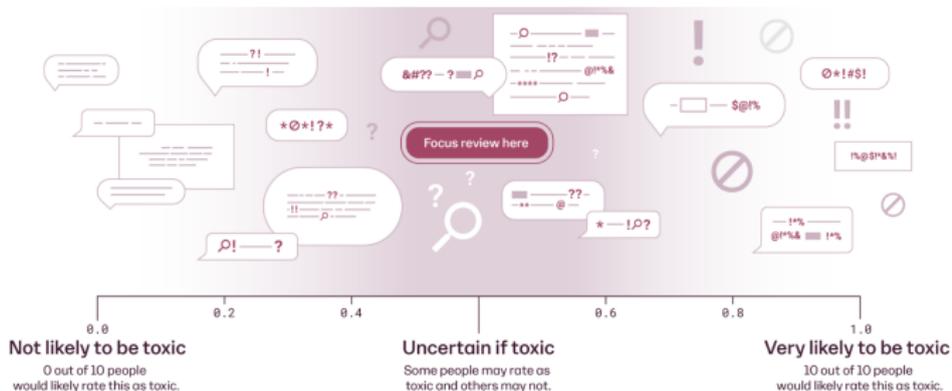
Outcomes and Controls

- Dependent Variables / Outcomes
 - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)



Outcomes and Controls

- Dependent Variables / Outcomes
 - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)



Outcomes and Controls

- Dependent Variables / Outcomes
 - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)

Toxicity Level	Description of level
Very Toxic	A comment that is very hateful, aggressive, disrespectful, or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.
Toxic	A comment that is rude, disrespectful, unreasonable, or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective.
Not Toxic	A neutral, civil, or even nice comment very unlikely to discourage the conversation.
I'm not sure	The comment could be interpreted as toxic depending on the context but you are not sure.

Outcomes and Controls

- Dependent Variables / Outcomes
 - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)

Category	Definition
Profanity/ Obscenity	Swear words, curse words, or other obscene or profane language.
Identity-based negativity	A negative, discriminatory, stereotype, or hateful comment against a group of people based on criteria including (but not limited to) race or ethnicity, religion, gender, nationality or citizenship, disability, age, or sexual orientation.
Insults	Inflammatory, insulting, or negative language towards a person or a group of people. Such comments are not necessarily identity specific.
Threatening	Language that is threatening or encouraging violence or harm, including self-harm.

Outcomes and Controls

- Dependent Variables / Outcomes
 - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)

COMMENT

You're a real idiot, you know that.

This comment is not in English or is not human-readable.

Rate the toxicity of this comment.

Very toxic: A comment that is very hateful, aggressive, disrespectful, or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.

Toxic: A comment that is rude, disrespectful, unreasonable, or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective.

- Very toxic
- Toxic
- Maybe, not sure
- Not Toxic

Does this comment contain obscene or profane language?

Profanity/obscenity: Swear words, curse words, or other obscene or profane language.

- Yes
- Maybe, not sure
- No

Does this comment contain identity-based negativity?

Identity-based negativity: A negative, discriminatory, stereotype, or hateful comment against a group of people based on criteria including (but not limited to) race or ethnicity, religion, gender, nationality or citizenship, disability, age, or sexual orientation.

- Yes
- Maybe, not sure
- No

Outcomes and Controls

- Dependent Variables / Outcomes
 - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)
 - TweetCount: $\log(\# \text{ of tweets in the Pre/Post-periods} + 1)$.

Outcomes and Controls

- Dependent Variables / Outcomes
 - Toxicity: mean(toxicity of the most toxic 10 percent of tweets)
 - TweetCount: $\log(\# \text{ of tweets in the Pre/Post-periods} + 1)$.
- Control Variables
 - Candidate Level
 - Overall Toxicity (Candidate toxicity bin)
 - Party ID, Sex, Incumbency, Twitter Account Type
 - Election Level
 - State, District, Type, District Competitiveness
 - Treatment Level
 - Sanctioning Timing and Group
 - Monitoring Date
 - Other Arm

Experimental Conditions

- 1 Top-Down **Monitoring** Condition
(Grossman and Hanlon, 2014; Grossman and Michelitch, 2018;
Nyhan and Reifler, 2015)
- 2 Bottom-up **Social Sanctioning** Condition
(Munger 2017)
- 3 Control Condition

Top-Down Monitoring Arm

*Dear [**\$Candidate Full Name**].*

We are two independent researchers at North Dakota State University. We are not affiliated with any partisan group in any way.

*We are writing to let you know we are conducting research on the use of toxic language on Twitter by candidates, specifically how use of such language affects election outcomes. We are monitoring your Twitter account [**@handle(s)**] and will compile your tweets that use toxic language. Just before the election, we will write a post on the Monkey Cage blog of the Washington Post that discusses our findings regarding patterns in the use of toxic language.*

Sincerely, Drs. Daniel Pemstein and Yunus Orhan

Bottom-Up Social Sanctioning Arm

Mention	Preamble (1 of 5)	Text (1 of 5)
	Good day!	Just remember that some of your constituents may be upset by toxic messages like this.
	Hi!	This toxicity is upsetting to some of your constituents.
@[Harasser Candidate]	Hey there.	Some of your constituents are going to be upset by this toxic message.
	Hi there.	That is a toxic thing to say, and it will push away some of your constituents.
	Hello!	Toxic messages like this will alienate some of your constituents.

Progressive Pictures & Banners

Brian Smith
 News junkie
 Beer drinker



Conservative Pictures & Banners



Sanctioning Example

Dr. Kelli Ward  @kelliwardaz · Oct 26, 2022 ...

1/2 People we send to ALL levels of govt should have the ability to do the job. These jobs require thinking, listening, reading, understanding, & communicating on many levels. If a person cannot do these things, they are unqualified for the job. Stop voting for people who can't.

👤 67 🔄 98 ❤️ 302 📤

Sanda Clark For Arizona, LD19 @SandaClark4AZ ...

Replying to @kelliwardaz

Hey, Kelli, you forgot to lie and cheat! That's your line of work! Hey, have you released your phone records to the feds yet? The wheels of justice are starting to roll!

8:56 AM · Oct 26, 2022

2 Retweets 7 Likes

🗨️ 🔄 ❤️ 📤

 Tweet your reply Reply

Dan Miller @DanMillerUSAYAY · Oct 26, 2022 ...

Replying to @SandaClark4AZ and @kelliwardaz

Hi! Toxic messages like this will alienate some of your constituents.

👤 🗨️ 🔄 ❤️ 3 📤

Sanctioning Example

Vanessa Enoch, Ph.D. @DrVEnoch · Oct 30, 2022

During our debate, my opponent @warrendavidson called for reinforcements, and now Russians are coming to his aid! Y'all take note... my opponent is ABSOLUTELY working for Putin! That's why they are defending him! secure.actblue.com/donate/vote-en... Donate to send him and the Russians packing.

Wyatt Reed @wyattreed13 · Oct 29, 2022

Russia state-affiliated media

5 of the top 65 deadliest cities in the US are in Ohio. It has low wages, twice the national rate of drug addiction and some of most overdoses in the country.

But none of this made Enoch run—the possibility that our much-needed money might go to Americans instead of Ukraine did. [twitter.com/DrVEnoch/statu...](https://twitter.com/DrVEnoch/status...)

Dave Williams @DaveWill89047170

Replying to @DrVEnoch

Hi! Toxic messages like this will alienate some of your constituents.

10:23 AM · Oct 31, 2022

2 Likes

Dr Dr Dr Eloi Morlock, JD, PhD, Time ... @jamesca... · Oct 31, 2022

Replying to @DaveWill89047170 and @DrVEnoch

Kudos for your exceptionally diplomatic phrasing.

Sanctioning Criteria

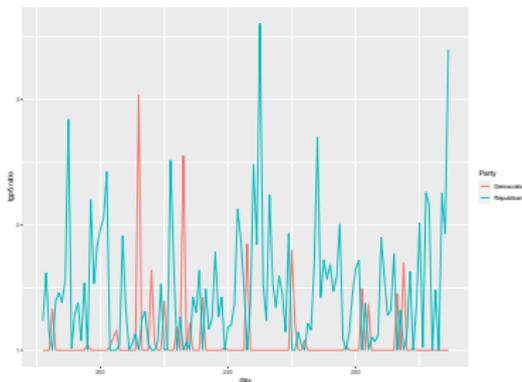
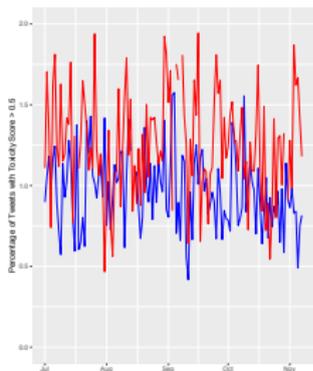
- The Google's Jigsaw Perspective (GJP) scores are imperfect **proxies** for sanctionable content.
- Hybrid Procedure: Combining machine coding and human discretion

Estimands

- Intention to treat (ITT) effect
 - Differences-in-differences (DiD) design
 - Pre/post treatment periods (1 week)
 - Quantity of interest is interaction between condition and period
- Total effect of treatment on the treated (TOT)
 - Sanctioning: only 121 actual sanction events
 - Monitoring: a handful of bounced emails
 - DiD, with instrumental variable (experimental condition)

Descriptive Findings

- 1 **Candidates** rarely engage in stark toxicity on Twitter
- 2 **Democrats** generate more toxic tweets
- 3 **Republicans'** tweets are more likely to be toxic than **Democrats'** tweets
- 4 **Men** are twice more likely than **women** to post toxic tweets



Top-Down Monitoring Intervention Results

	ITT				TOT			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.097*** (0.003)	-29.8 (18.8)	0.135*** (0.004)	-51.4** (24.2)	0.097*** (0.003)	-30.3 (18.9)	0.135*** (0.004)	-52.0** (24.2)
Treatment	0.011** (0.005)	0.011*** (0.004)	0.013** (0.006)	0.014*** (0.005)	0.011** (0.005)	0.012*** (0.004)	0.014** (0.006)	0.015*** (0.005)
Period	0.015*** (0.005)	0.015*** (0.004)	0.017*** (0.006)	0.017*** (0.005)	0.015*** (0.005)	0.015*** (0.004)	0.017*** (0.006)	0.017*** (0.005)
Treatment × Period	-0.014** (0.007)	-0.014*** (0.005)	-0.015* (0.008)	-0.017** (0.007)	-0.015** (0.007)	-0.015** (0.006)	-0.016* (0.009)	-0.018** (0.007)
Controls	X	✓	X	✓	X	✓	X	✓
0-Tweet	0	0	NA	NA	0	0	NA	NA
R ²	0.002	0.425	0.002	0.391	0.002	0.425	0.002	0.391
Observations	7,120	7,120	5,223	5,223	7,120	7,120	5,223	5,223

- Monitoring **reduces** P(toxic) by 1.5 points, on average
- This represents a 15% reduction for a typical candidate

Bottom-Up Social Sanctioning Intervention Results

	ITT				TOT			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.105*** (0.003)	0.015 (0.011)	0.141*** (0.004)	0.043*** (0.015)	0.105*** (0.003)	0.013 (0.011)	0.141*** (0.004)	0.048*** (0.015)
Treatment	-0.003 (0.005)	-0.006 (0.004)	0.004 (0.006)	0.003 (0.005)	-0.050 (0.072)	-0.058 (0.046)	0.061 (0.089)	0.031 (0.059)
Period	-0.003 (0.005)	-0.003 (0.004)	-0.002 (0.006)	-0.002 (0.005)	-0.003 (0.005)	-0.001 (0.003)	-0.002 (0.006)	-0.003 (0.004)
Treatment × Period	0.011 (0.007)	0.011** (0.005)	0.003 (0.009)	0.004 (0.007)	0.163 (0.102)	0.104** (0.049)	0.047 (0.126)	0.084 (0.061)
Controls	X	✓	X	✓	X	✓	X	✓
0-Tweet	0	0	NA	NA	0	0	NA	NA
R ²	0.0005	0.401	0.0004	0.371	0.0005	0.401	0.0004	0.371
Observations	7,120	7,120	5,184	5,184	7,120	7,120	5,184	5,184

- Sanctioning **increases** P(toxic) by 10 points, on average
- This represents a 100% increase for a typical candidate

Chilling Effect?

	ITT		TOT	
	(1)	(2)	(3)	(4)
Constant	1.63*** (0.029)	-168.4 (158.2)	1.63*** (0.029)	-164.2 (158.3)
Treatment	-0.013 (0.042)	-0.012 (0.031)	-0.014 (0.045)	-0.013 (0.034)
Period	0.077* (0.042)	0.077** (0.031)	0.077* (0.042)	0.077** (0.031)
Treatment × Period	-0.051 (0.059)	-0.051 (0.044)	-0.055 (0.064)	-0.055 (0.047)
Controls	X	✓	X	✓
R ²	0.0008	0.447	0.0008	0.447
Observations	7,120	7,120	7,120	7,120

- **No evidence** that candidates tweet less when monitored

Key Takeaways

- 1 Top-down monitoring **REDUCES** toxicity among elites
- 2 Bottom-up sanctioning **INCREASES** toxicity among elites
- 3 Top-down monitoring **HAS NO EFFECT** on candidate tweet volume

Key Takeaways

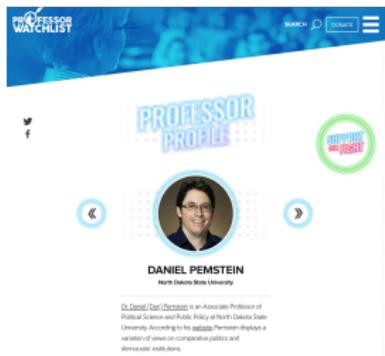
- 1 Top-down monitoring **REDUCES** toxicity among elites
- 2 Bottom-up sanctioning **INCREASES** toxicity among elites
- 3 Top-down monitoring **HAS NO EFFECT** on candidate tweet volume

What's next?

Building our own metaketa
Better monitoring interventions
Custom toxicity measures

Key Takeaways

- 1 Top-down monitoring **REDUCES** toxicity among elites
- 2 Bottom-up sanctioning **INCREASES** toxicity among elites
- 3 Top-down monitoring **HAS NO EFFECT** on candidate tweet volume



What's next?
Building our own metaketa
Better monitoring interventions
Custom toxicity measures



Thanks!

RAs:

William Carney, Madison Delorme, Antony Ibrahim, Evan Jones,
Maguire Martin, Ruben Orozco, Jonathan Ross, Rohan Tapiawala

Collaborators:



Funding:

