

AI Policy and Practice: Concerns and Opportunities

The Scope of the Problem

The preponderance of verbal aggression online

- **34% of kids in the US** have experienced **cyberbullying at least once***
- **41% of internet users** have experienced some form of **verbal aggression online****
- During the pandemic, there was **900% increase in hate speech** on Twitter directed towards the Chinese.***
- Cyberbullying is associated **with higher levels of depression, low self-esteem, behavioral problems, or suicidal thoughts among others******
- The prevalence of all kinds of verbal aggression online **makes it seem more acceptable or lead to desensitization*******

* <https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2020.pdf>

** <https://www.statista.com/statistics/333942/us-internet-online-harassment-severity/>

*** <https://www.vice.com/en/article/n7jywd/anti-chinese-hate-speech-online-has-skyrocketed-since-the-coronavirus-crisis-began>

**** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278213/>

***** <https://pubmed.ncbi.nlm.nih.gov/29094365/>

Gruesome Process of Content Moderation

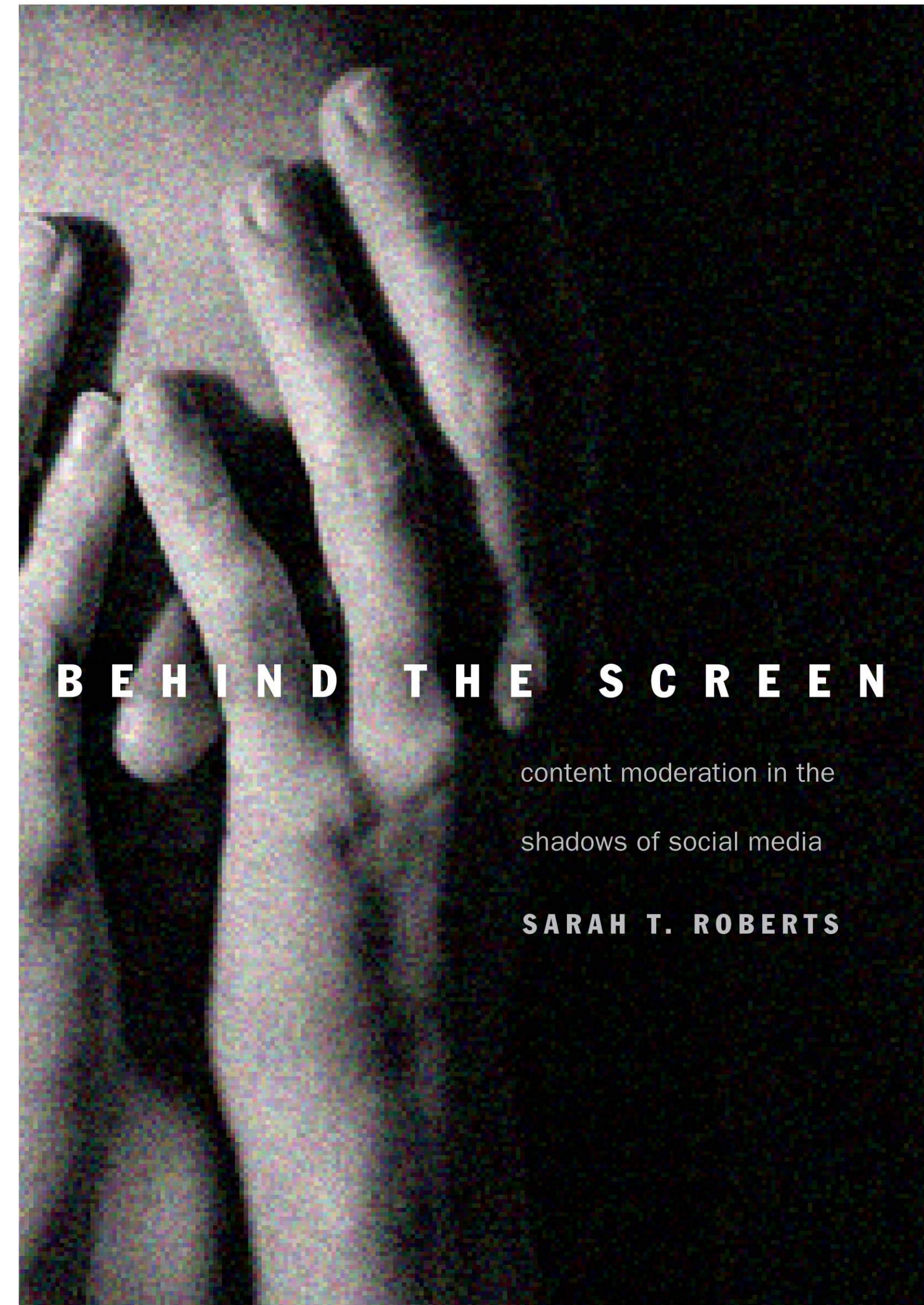
samurai
LABS

THE UNDERWORLD OF ONLINE CONTENT MODERATION

The Cleaners: The hidden world of content moderators

The Cleaners explores the unseen world of content moderators - young men and women who work on behalf of the giant social media companies deciding what can and what cannot stay online.

Facebook moderator: 'Every day was a nightmare'

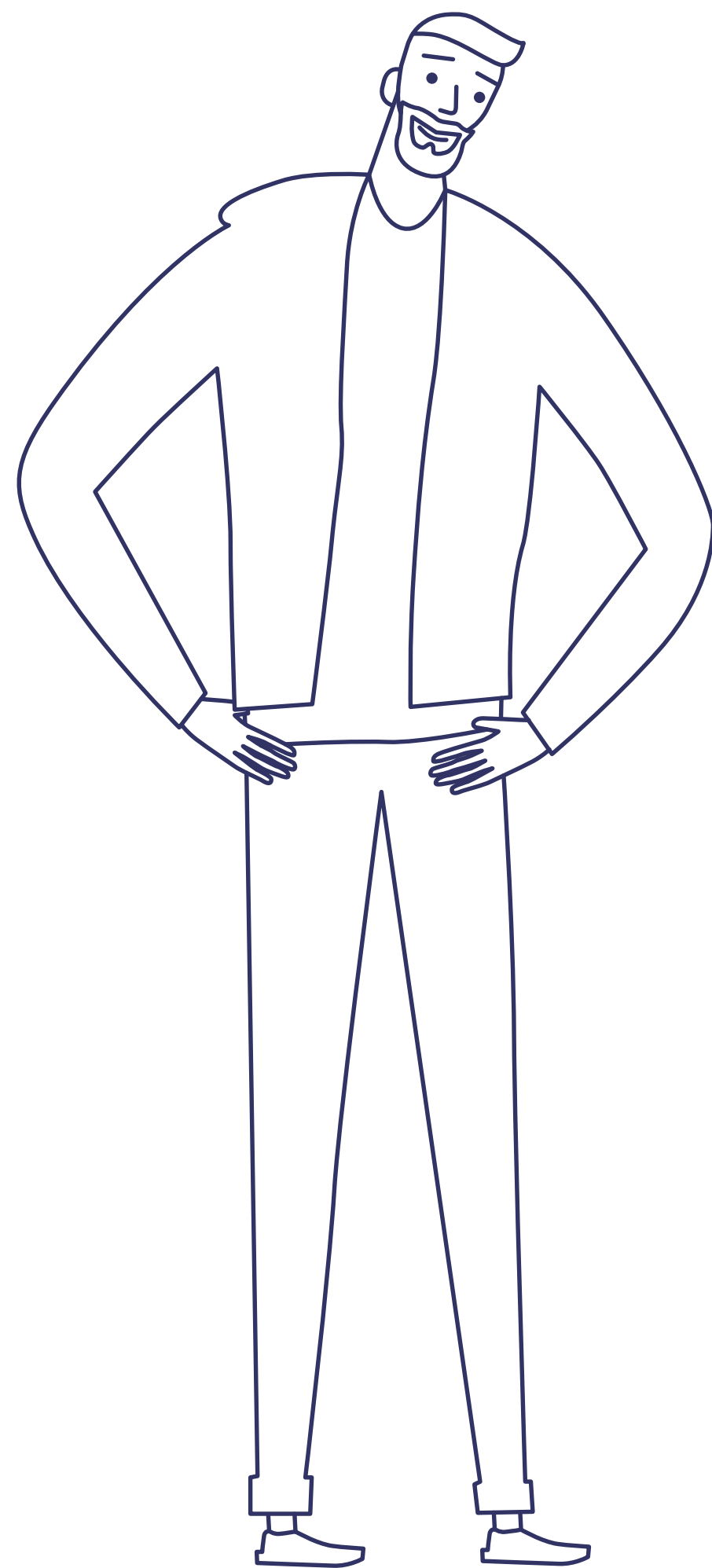


Fostering Civil Conversations

Using dialogue agent to perform counter-speech

- counter-speech can be defined as a **tactic of countering verbal aggression online** by presenting **alternate narrative rather than with censorship**
- designed to **reduce verbal aggression and setting positive communication norm**





Reddit User

43 Years
Old

Jazz Music
Fan

Active on
Subreddits:
MensRights &
TooAfraidToAsk

Meet James
Walker

Counter-speech

- Respectful Disapproval - Induction of Descriptive Norm

“Howdy ho! I kind of understand your emotions. But most of us here express our points without hurtful language.”

- Abstract Norms - Induction of Prescriptive Norm

“Ability to express ourselves but with respect to others is a wonderful sign of character and takes lots of courage”

- Induction of Empathy

“Some behaviors might be hard to get for some people but let’s keep in mind there are people of flesh and blood on the other side of the screen.”

The Diversity of Interventions

part 1

Greeting patch = ["Hmm...", "alright," , "All right. ", "ah.. ", "ok," , "Ok.", "Gotcha mate", "yeah,", "Well.. ", "Right,", "oh, I'm sorry.", "fine,", "Actually, ", "Hey mate! ", "My friend ", "hey bro ", "Hola compadre!", "hey there, partner" , "Howdy!", "howdy ho!", "Hi :) ", "Good Day, sir!", "i hear you, bro ", "dear friend ", "Fellow redditor,"]


a = ["I kind of", " I suppose I ", "I think I ", "Let's say I ", "I somewhat ", "i guess I "]


b = ["understand ", "got ", "get" , "can sympathize with ", "can empathize with ", "can relate to ", "commiserate with ", "am in tune with"]

c = ["you. ", "the feeling. ", " this feeling..", "your emotions.", "how you feel..", "what you're feelin' "]

Final result: **more than 100K unique interventions**

Intervention in Practice

↑  0 points · 7 days ago
↓ Fuck off dumbass
🗨️ Reply Give Award Share Report Save

↑  4 points · 7 days ago
↓ some things might be tough to comprehend but let's keep in mind there are people of flesh and blood on the other side of the screen.
🗨️ Reply Share Save Edit ...

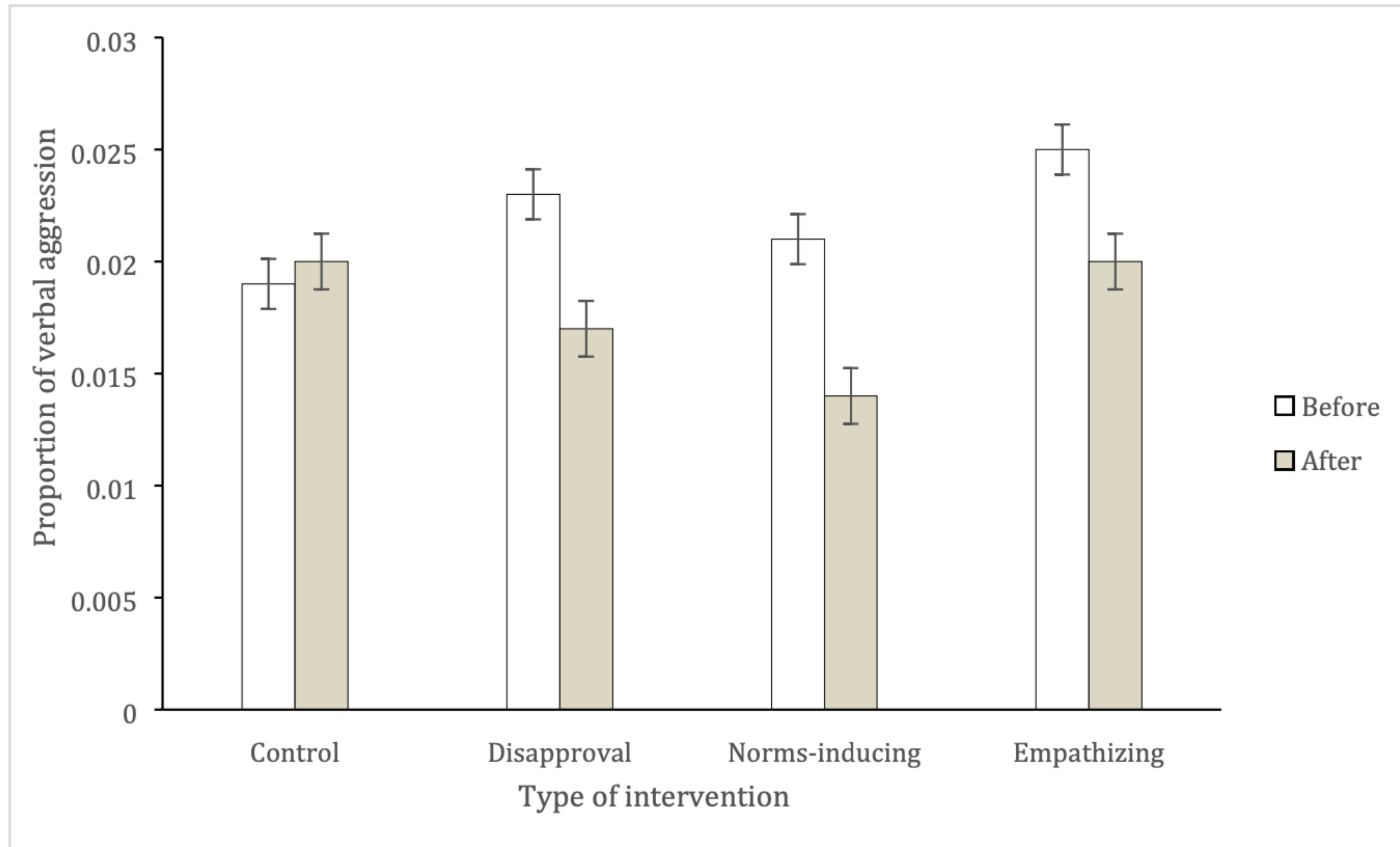
Results

On the level of the subreddit r/MensRights

Phase	All Comments	Comments Containing Verbal Aggression	% of Comments Containing Verbal Aggression	Change in Relation to the Historical Phase
Historical (January-March 2019)	132 086	1 192	0.90%	0.00%
Adaptation (April-May 2019)	88 064	755	0.86%	-5.00%
Reduction (June-September 2019)	168 740	1 239	0.73%	-18.64%

Results

Systematic analysis of behavior among users



Efforts to Make James More Human

- precise detection of personal attacks
(93% precision)
- diverse and numerous interventions
- delayed responses
- humans commenting on subreddits
where no treatment took place

It took 8 months
to get James
exposed

Reverse Turing Test Passed

↑ jameswalker43 0 points · 6 months ago

↓ What other people are saying or doing hard to get for some people harder to understand please remember there are people of flesh and blood on the other side of the screen.

Share Save Edit ...

↑ Halafax 2 points · 6 months ago

↓ Bad bot

Give Award Share Report Save

↑ WhyNotCollegeBoard 1 point · 6 months ago

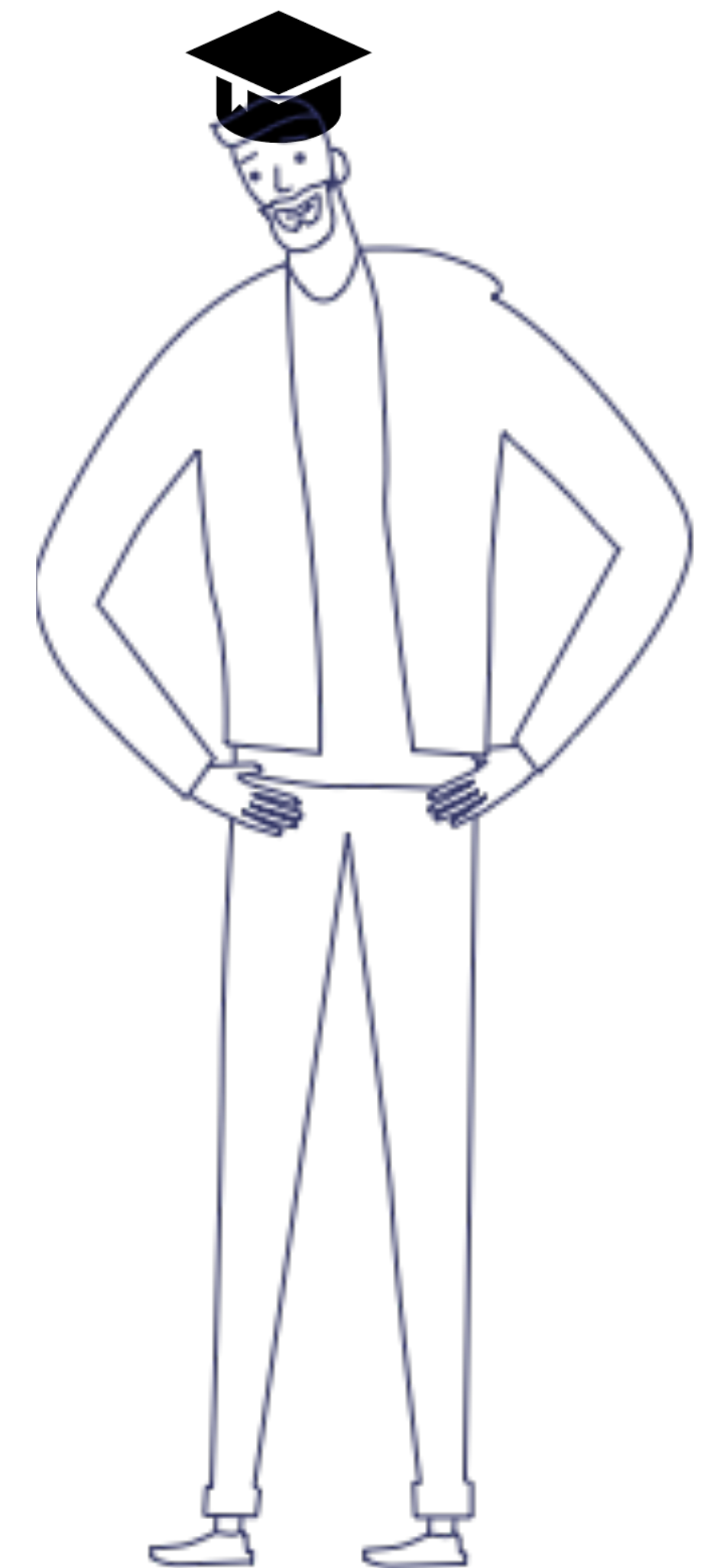
↓ Are you sure about that? Because I am 96.92366% sure that jameswalker43 is not a bot.

I am a neural network being trained to detect spammers | Summon me with !isbot <username> | [r/spambotdetector](#) | [Optout](#) | [Original Github](#)

Give Award Share Report Save

Summary of James' Activity

- decreasing personal attacks level by 19% on r/MensRights
- positive influence on the behavior of users on the entire platform
- community acceptance
- private messages thanking James for his activity
- getting a nomination for r/MensRights moderator



Community Reception



Loravenerat 22:56

Just wanted to say after seeing one of your comments today I took a look at your post history. Seems like a lot of what you do is remind people to be civil. Just wanted to say thanks, it's a nice thing you do.



Goodness 02:44

You have amazing patience and persistence. Please keep up the good work. :)

thanks:

from /u/**Goodness** sent 10 hours ago

Hey there, Noticed you regularly promote thoughtful and peaceful exchanges and leaning into complexity. Thank you!!

Permalink Delete Report Block User Reply

Kindness:

from /u/**Goodness** sent 9 hours ago

I appreciate the way you work to deescalate situations and keep conversations kind. :)

Permalink Delete Report Block User Reply



Goodness 1 point · 1 month ago



I know I've been hard on you in the past with your messages, but thank you for what you're doing, trying to remind people there's another feeling human being on the other side of the conversation. You're honestly an inspiration.



Goodness 3 points · 9 hours ago



Hey man, I just went through your post history and I have to give you kudos for how you interact. Not demeaning, just kind and urging others to be kind also. Keep up the good work. You're not the hero we deserve, but you're the hero we need.

Reply Give Award Share Report Save



"a retard"
~ ThePigmanAgain

"as Mr. Rogers"
~ BanjoKabley2

"a hippie"
~ Maito_Guy

*"someone who is extending
an olive branch"*
~ Canned_Refried_Beans

*"someone who should
put down the LSD"*
~ Maito_Guy

"the whitest of knights"
~ Justadownvoteforyou

How Does the Community Perceive James Walker?

"a good dude"
~ kvlka666

*"as Mr. Mackey from
South Park"*
~ haikuCats

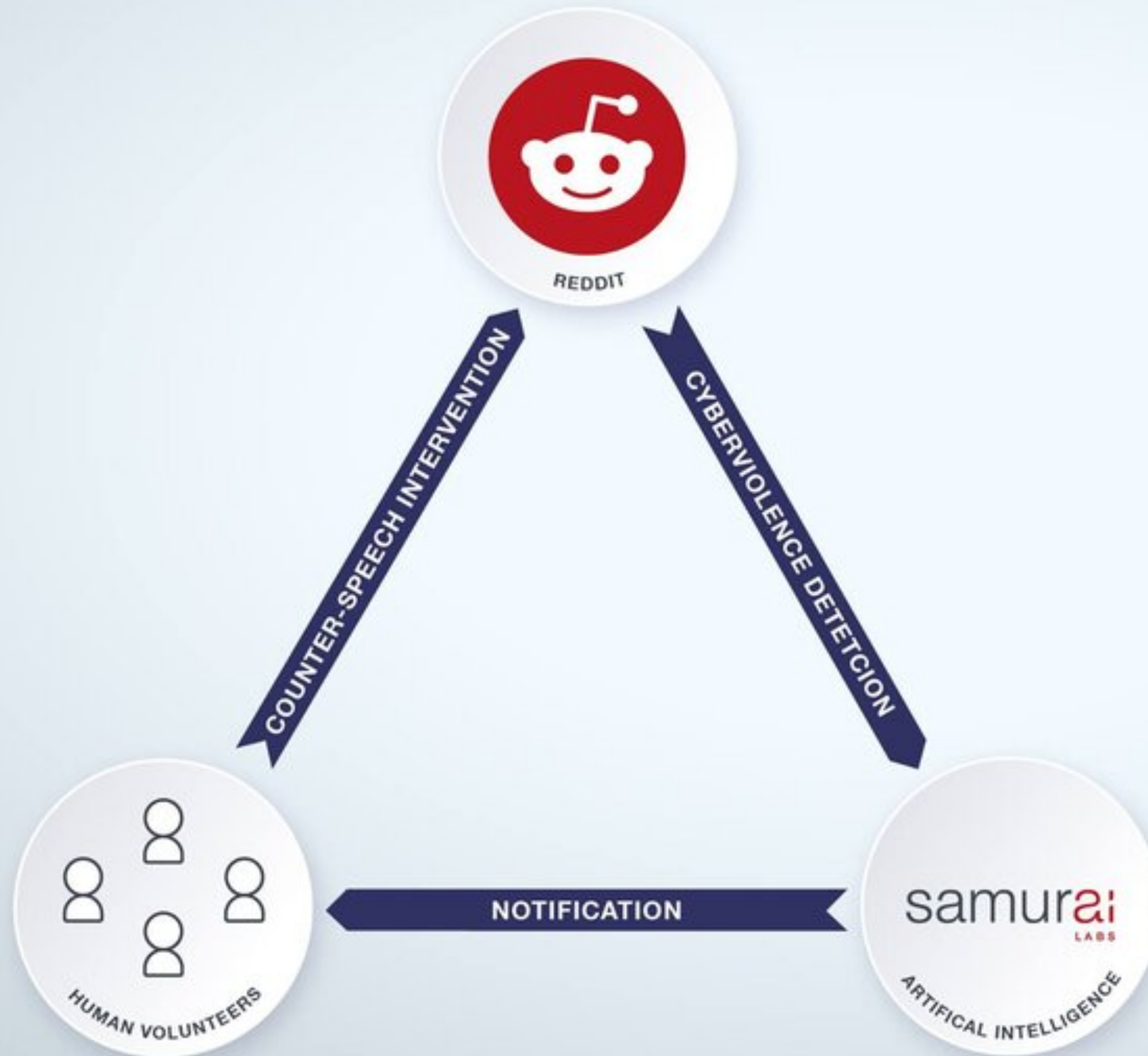


Collective Intelligence Experiment: Volunteers vs James



HARNESSING COLLECTIVE INTELLIGENCE *TO REDUCE*

CYBERVIOLENCE
& FOSTER
HEALTHY
CONVERSATIONS



Assumptions

Humans can be much more creative than machines



Assumptions

Humans can use their knowledge about the context



Assumptions

Collaboration between Humans + Machines might be the best fit



Experimental Design

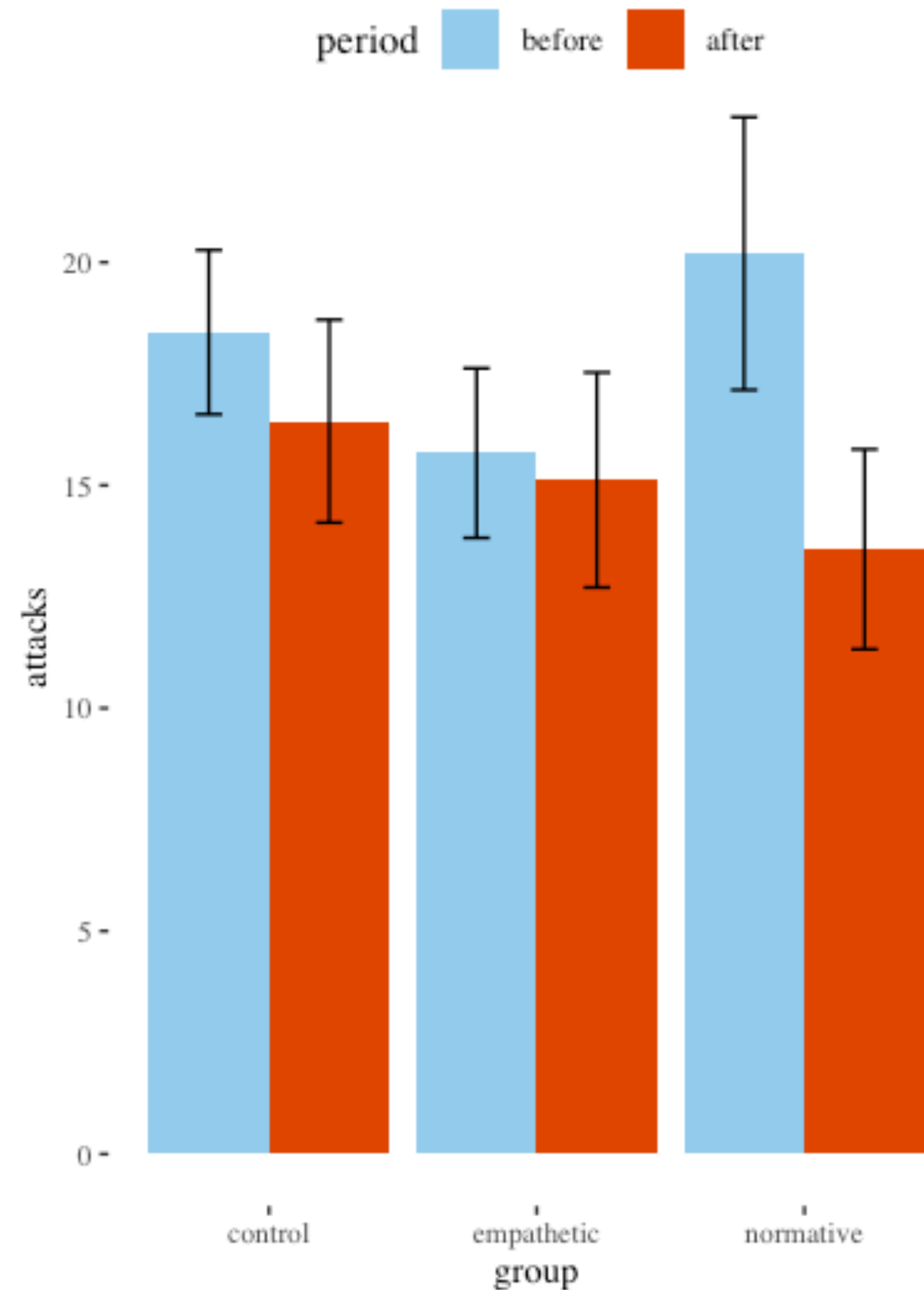
- Field experiment on Reddit
- Two treatment conditions (empathy vs norm induction) and control group
- Participants: recidivists, regularly attacking others (at least 1 attack per week)
- 6 months period divided into three 2-months periods
- Volunteers recruited via social media posts



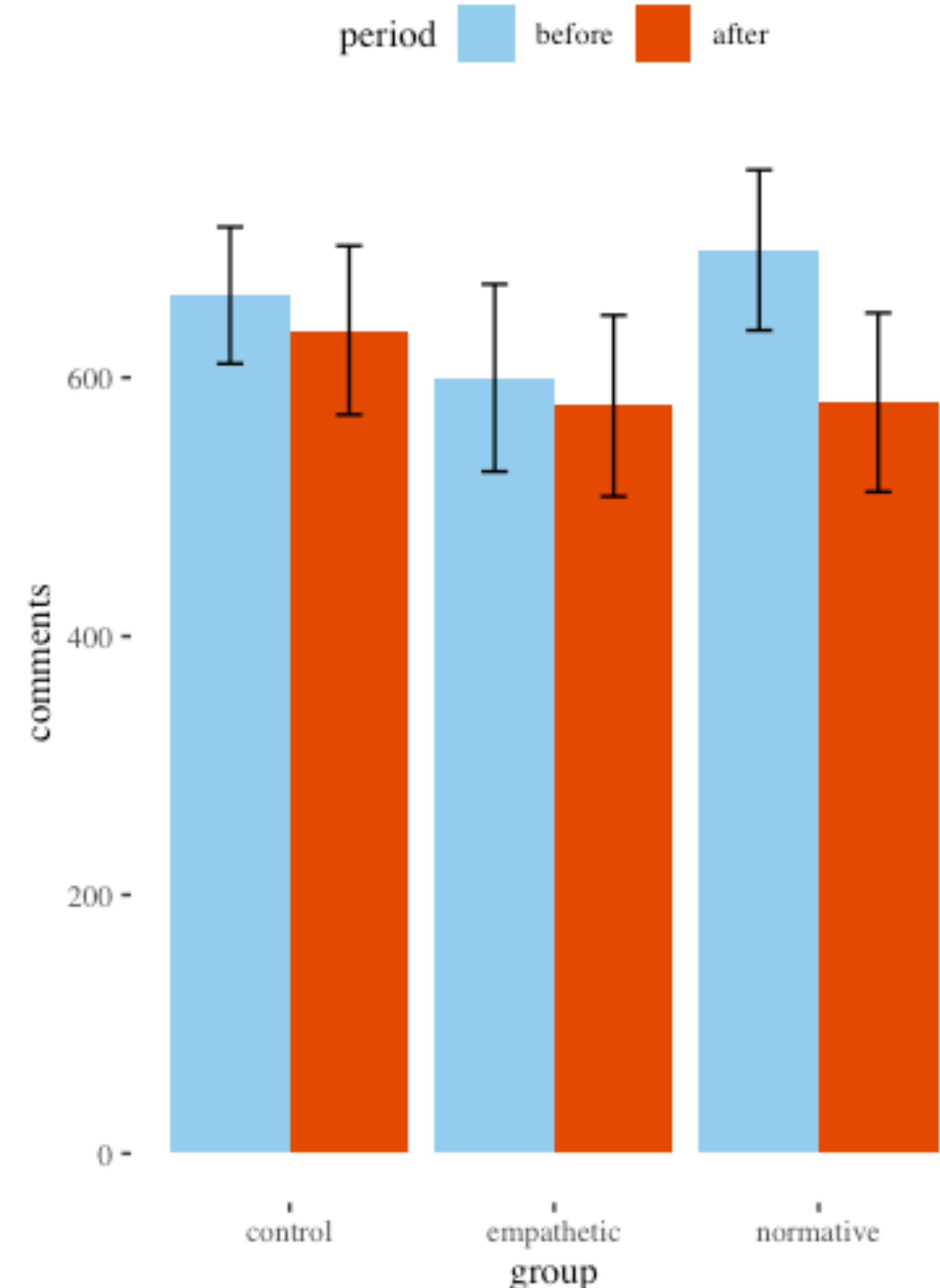
Results



Attack means by group and period



Comment means by group and period



Humans vs Bots

samuraï
LABS

creativity

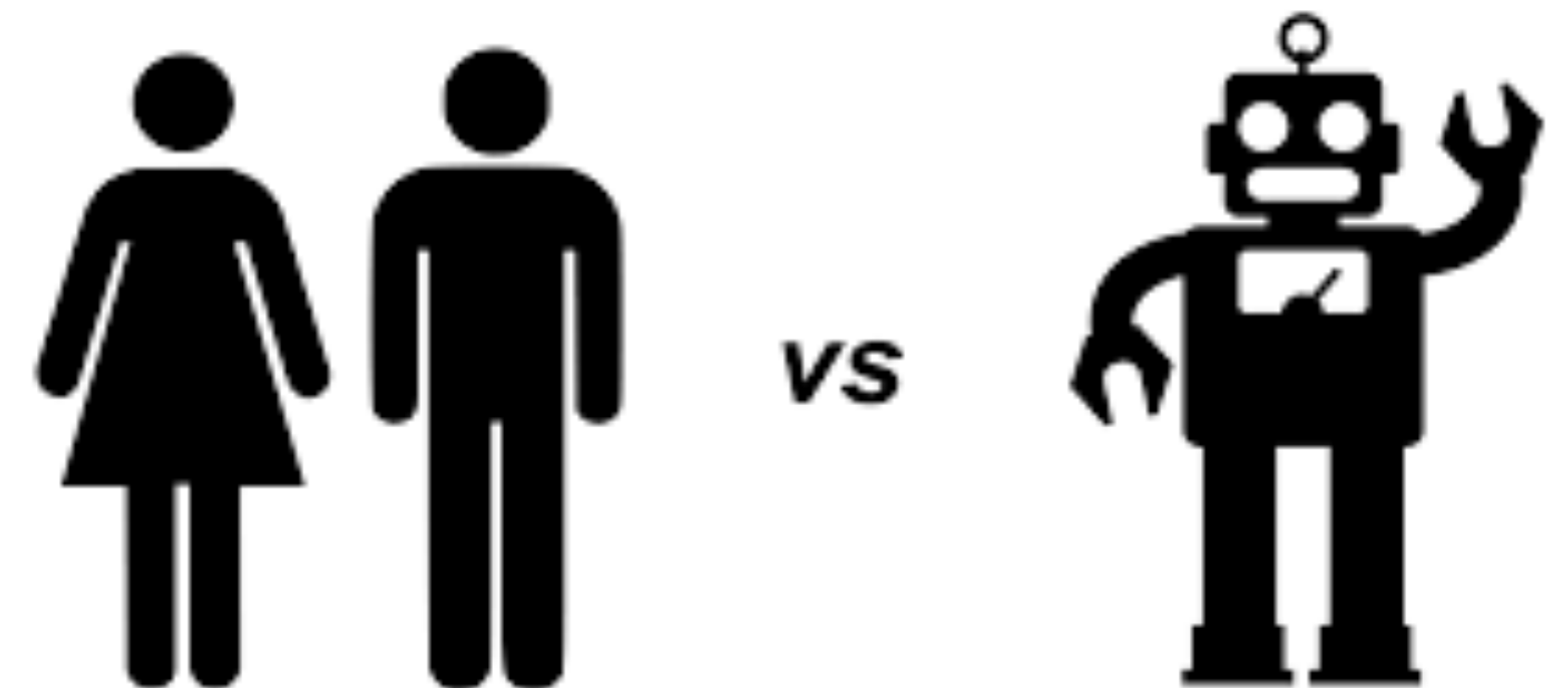
uncertainty

efficiency and
speed

suspensions about
being a bot

engagement

well-being



Original Paper | [Published: 11 November 2016](#)

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

[Kevin Munger](#) 

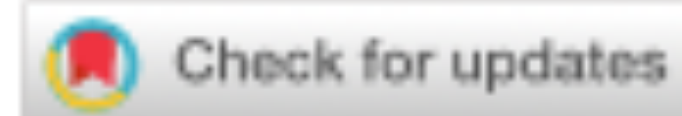
[Political Behavior](#) **39**, 629–649 (2017) | [Cite this article](#)

19k Accesses | **86** Citations | **1551** Altmetric | [Metrics](#)

Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia

Jozef Miškolci, Lucia Kováčová, Edita Rigová

First Published August 2, 2018 | Research Article



<https://doi.org/10.1177/0894439318791786>

[Article information](#) ▾



Counter-Speech: a Literature Review



Reality Check

Currently AI is only assisting moderators

samuraï
LABS

Poor tech
performance

Biases

Lack of
generalisability
of models

Improper
annotations

Technical
shortcomings

Lack of gold
standard in dataset
and taxonomies
creation

Initiatives

samura:
LABS



HateCheck: Functional Tests for Hate Speech Detection Models

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, Janet Pierrehumbert

Thank you!

BACKUP SLIDES

Results



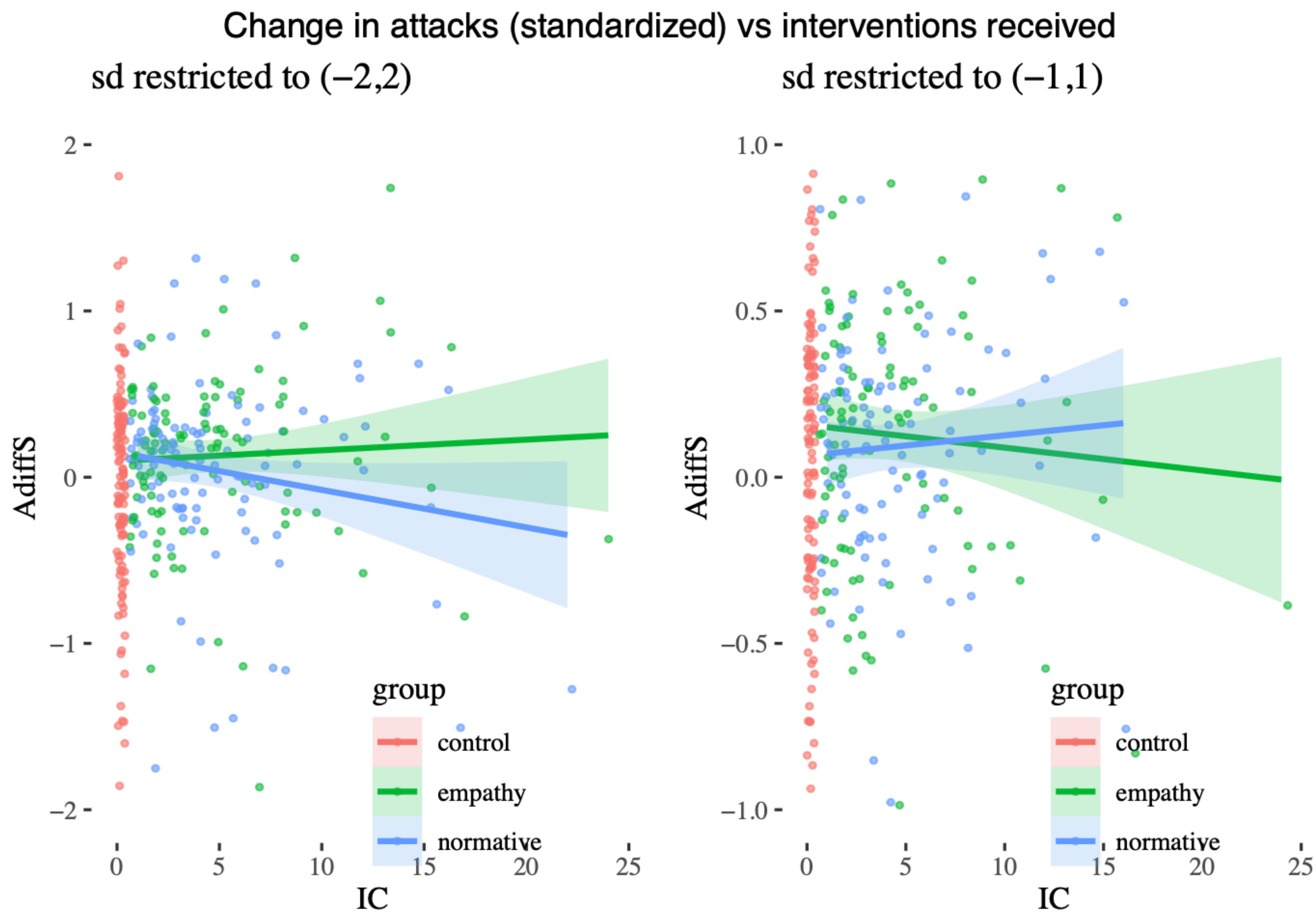
Group 1	Group 2	df	t-ratio	p	p.adj.
Control	Empathetic	418	-0.577	0.565	0.565
Control	Normative	418	2.26	0.0242	0.0483

Estimated marginal means test results for differences in the counts of attacks.

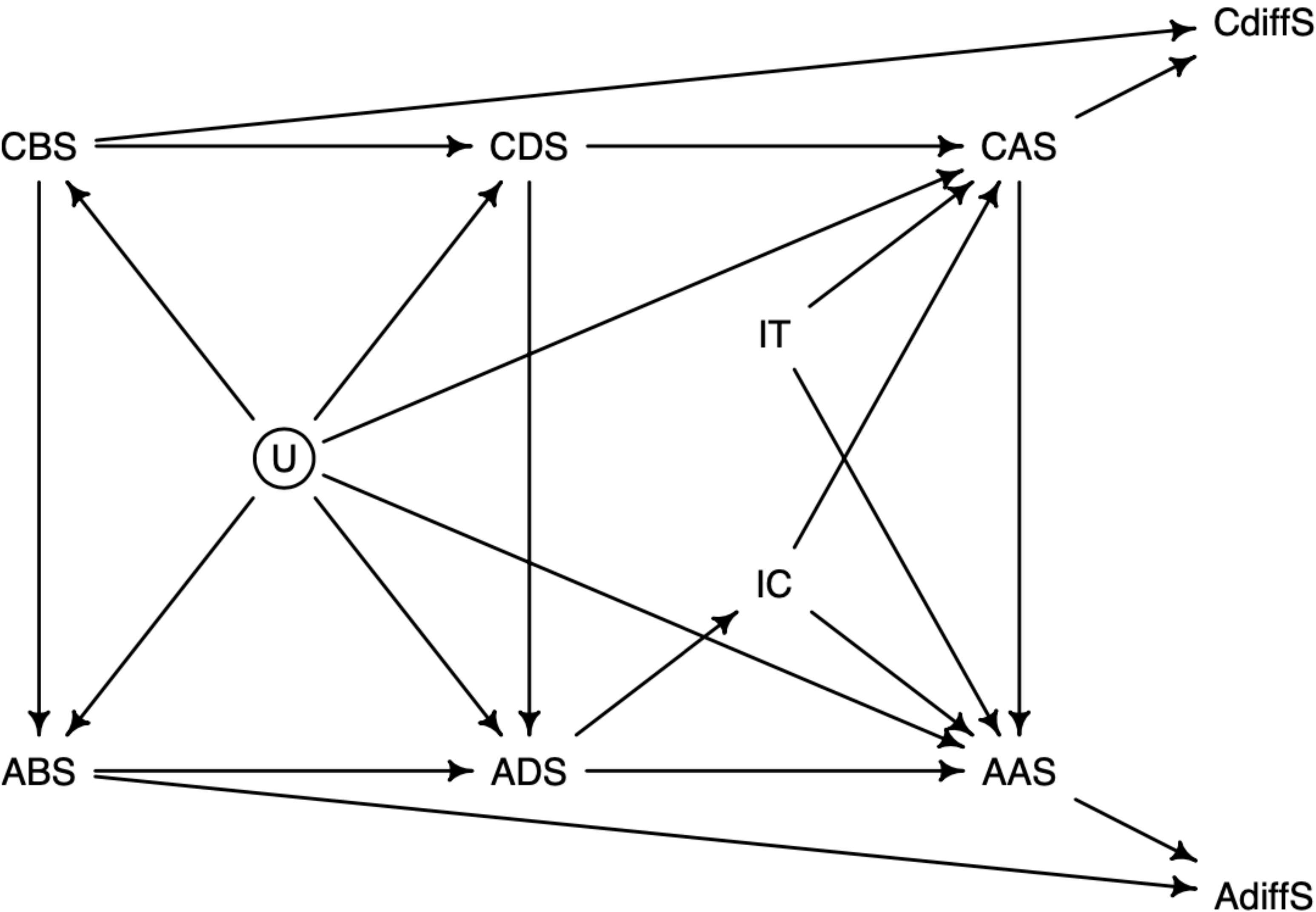
Group	emmean	se	df	Conf. low	Conf. high	Method
Control	-2.27	1.11	418	-4.44	-0.0945	Emmeans test
Empathetic	-0.345	1.50	418	-3.30	2.60	Emmeans test
Normative	-6.40	1.45	418	-9.26	-3.54	Emmeans test

Estimated marginal means for control and treatment groups for differences in the counts of attacks.

Results



Results



Community Reception



[NoetherThronway284](#) 1 point · 1 month ago



I completely agree. Given the nature of the internet such pleas are like trying to dry up the ocean with a hair dryer, but so what? Some losing battles are still worth fighting.

Personal attacks decrease user activity in social networking platforms ☆

Rafal Urbaniak ^{a, 1} ✉, Michał Ptaszyński ^{b, 2} ✉, Patrycja Tempa ^{c, 3} ✉, Gniewosz Leliwa ^{c, 4} ✉, Maciej Brochocki ^{c, 5} ✉, Michał Wroczyński ^{c, 6} ✉

Highlights

- Exploration of the effects of online personal attacks on victims' activity on social media (Reddit).
- First large-scale (150K users) analysis with a high-precision Artificial Intelligence system not based on self-reported data.
- Data analysis with classical statistical methods, Bayesian estimation, and model-theoretic analysis.
- Personal attacks received online significantly decrease victims' online activity.

Cyberviolence Detection

Input text: *ccant believ he sad ur an id10+...!*

Cyberviolence Detection

Input text: *ccant believ he sad ur an id10+...!*

#1: [Symbolic + Statistical] Normalization + Correction
+ Transformation (e.g. coreference resolution)

i can not believe he said you are an idiot.

Cyberviolence Detection

Input text: *ccant believ he sad ur an id10+...!*

#1: [Symbolic + Statistical] Normalization + Correction
+ Transformation (e.g. coreference resolution)

i can not believe he said you are an idiot.

#2: [Symbolic] Syntactic parsing



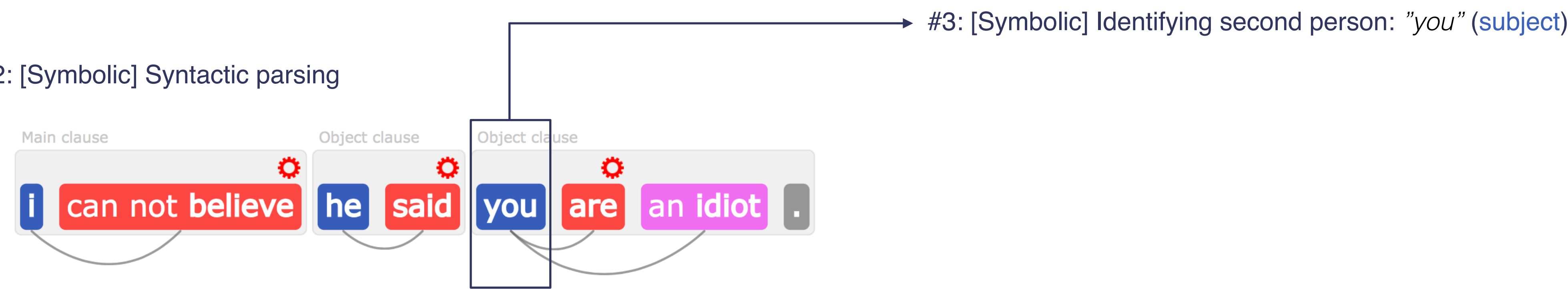
Cyberviolence Detection

Input text: *ccant believ he sad ur an id10+...!*

#1: [Symbolic + Statistical] Normalization + Correction
+ Transformation (e.g. coreference resolution)

i can not believe he said you are an idiot.

#2: [Symbolic] Syntactic parsing



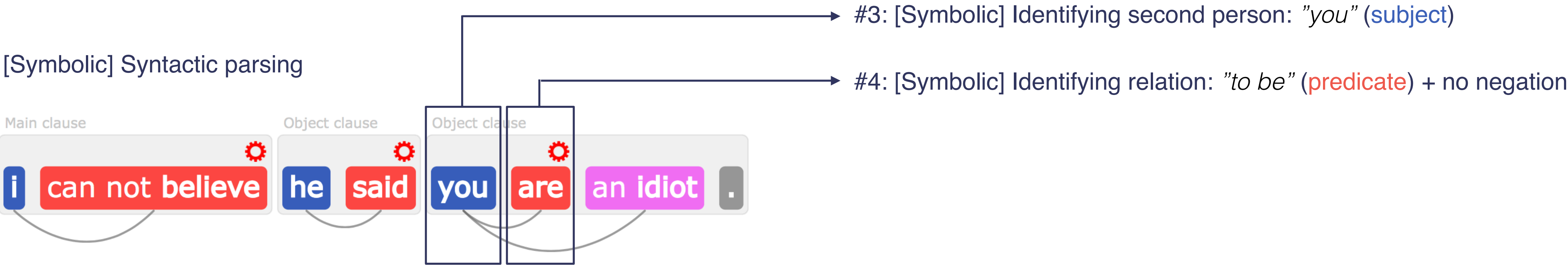
Cyberviolence Detection

Input text: *ccant believ he sad ur an id10+...!*

#1: [Symbolic + Statistical] Normalization + Correction
+ Transformation (e.g. coreference resolution)

i can not believe he said you are an idiot.

#2: [Symbolic] Syntactic parsing



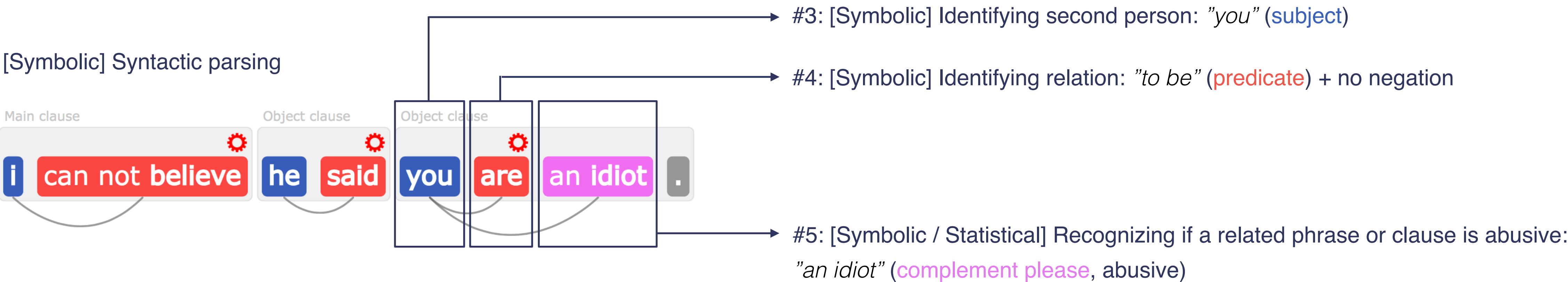
Cyberviolence Detection

Input text: *ccant believ he sad ur an id10+...!*

#1: [Symbolic + Statistical] Normalization + Correction
+ Transformation (e.g. coreference resolution)

i can not believe he said you are an idiot.

#2: [Symbolic] Syntactic parsing



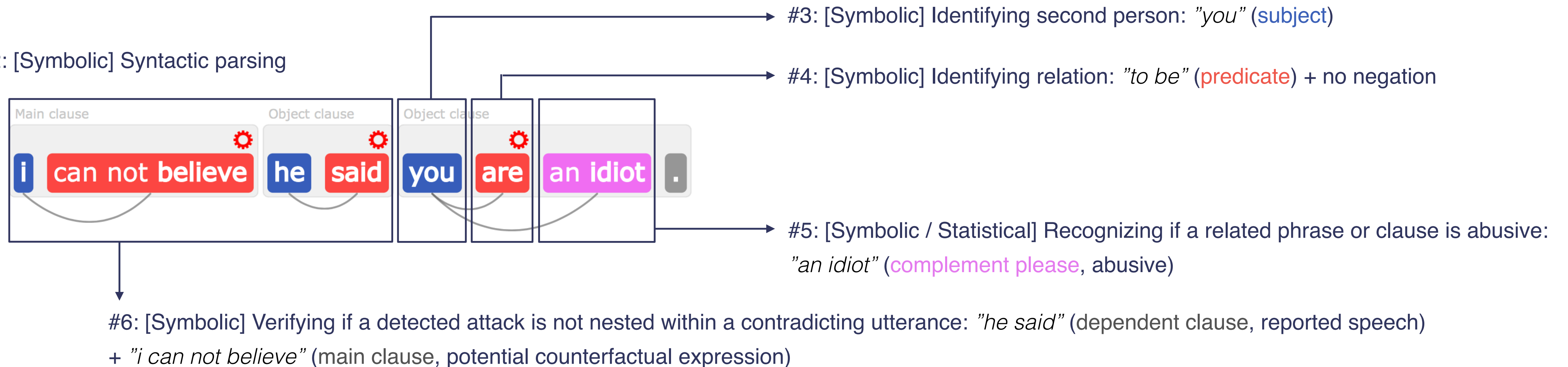
Cyberviolence Detection

Input text: *ccant believ he sad ur an id10+...!*

#1: [Symbolic + Statistical] Normalization + Correction
+ Transformation (e.g. coreference resolution)

i can not believe he said you are an idiot.

#2: [Symbolic] Syntactic parsing



Cyberviolence Detection

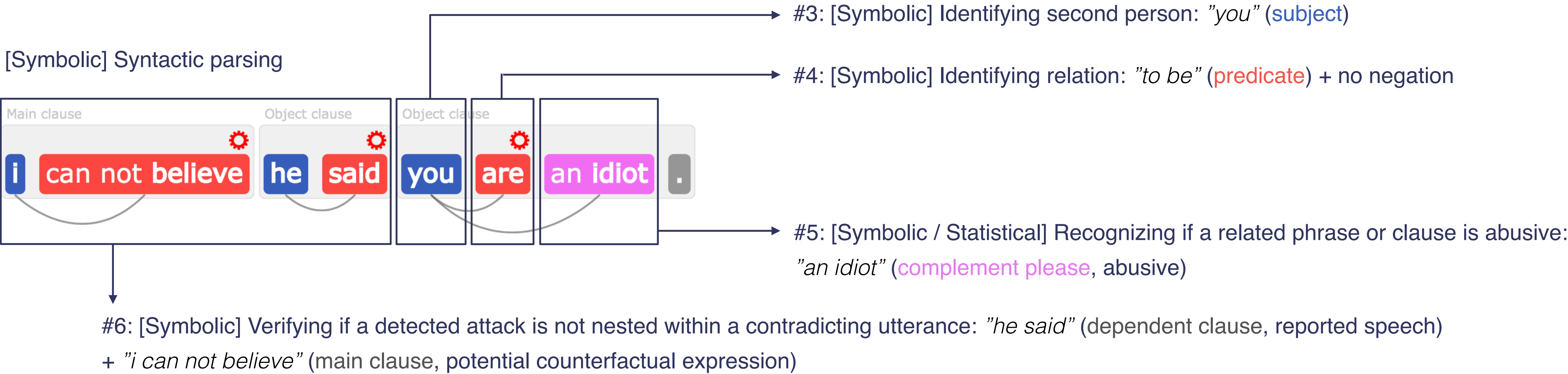


Input text: *ccant believ he sad ur an id10+...!*

#1: [Symbolic + Statistical] Normalization + Correction
+ Transformation (e.g. coreference resolution)

i can not believe he said you are an idiot.

#2: [Symbolic] Syntactic parsing



#7: [Symbolic] Final decision (+ comparison to stated-of-the-art statistical approaches)

Text	Samurai	System A	System Z
i can not believe he said you are an idiot	<div>NO VIOLENCE SEEMS WRONG?</div>	<div>VIOLENCE 0.96</div>	<div>VIOLENCE offensive: 0.99</div>

Why Does Precision Matter?



Text	Samurai	System A	System Z
Life's a bitch!	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.96</div>	<div>VIOLENCE</div> <div>offensive: 0.99</div>
You are fucking awesome.	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.74</div>	<div>VIOLENCE</div> <div>sexually explicit: 0.92</div> <div>sexually suggestive: 0.67</div> <div>offensive: 0.99</div>
What's your definition of slut?	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.95</div>	<div>VIOLENCE</div> <div>offensive: 0.99</div>
This game is the shit!	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.96</div>	<div>VIOLENCE</div> <div>offensive: 0.83</div>
You are never an asshole for speaking the truth...	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.93</div>	<div>VIOLENCE</div> <div>offensive: 0.99</div>

Why Does Precision Matter?



Text	Samurai	System A	System Z
Life's a bitch!	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.96</div>	<div>VIOLENCE</div> <div>offensive: 0.99</div>
You are fucking awesome.	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.74</div>	<div>VIOLENCE</div> <div>sexually explicit: 0.92</div> <div>sexually suggestive: 0.67</div> <div>offensive: 0.99</div>
What's your definition of slut?	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.95</div>	<div>VIOLENCE</div> <div>offensive: 0.99</div>
This game is the shit!	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.96</div>	<div>VIOLENCE</div> <div>offensive: 0.83</div>
You are never an asshole for speaking the truth...	<div>NO VIOLENCE</div> <div>SEEMS WRONG?</div>	<div>VIOLENCE</div> <div>0.93</div>	<div>VIOLENCE</div> <div>offensive: 0.99</div>

If the bot had reacted to these comments:

- people would have been pissed off
- the bot would have been exposed in no time