Understanding Online Counterspeech

Identity, Motivations, and Barriers in Responding to Hate

Kaike Ping

CSCW 2025

SAIL-Lab

Dr. Eugenia Rho
(Director)

Current

Xiaohan Ding
(PhD Student)

Kaike Ping
(PhD Student)

Buse Carik
(PhD Student)

Lance Wilhelm
(PhD Student)

Taufiq Daryanto
(PhD Student)

Kirk Knutsen
(PhD Student)

Caleb Wohn
(PhD Student)

Alumni

🎓 Sophia Stil
(MS Student)

🎓 Anisha Kumar
(MS Student)
Booz Allen Hamilton

🎓 Uma Gunturi
(MS Student)
Machine Learning Engineer, IBM

🎓 Rohan Leeha
(MS Student)
Researcher, MIT Lincoln Laboratory

Society + Ai & Language
Research Lab

VIRGINIA TECH.

The Challenge of Online Hate

Hate speech is a pervasive problem on social media.

Counterspeech—directly responding to hate—is a key strategy.

little is known about the people who write counterspeech.

Our Research Questions

① Motivates & Barriers

② Identity & Experience

Our Method: A Large-Scale Survey

458 English-speaking U.S. adults

Each participant saw 3 hate speech examples from 900 posts.

Topic Categories

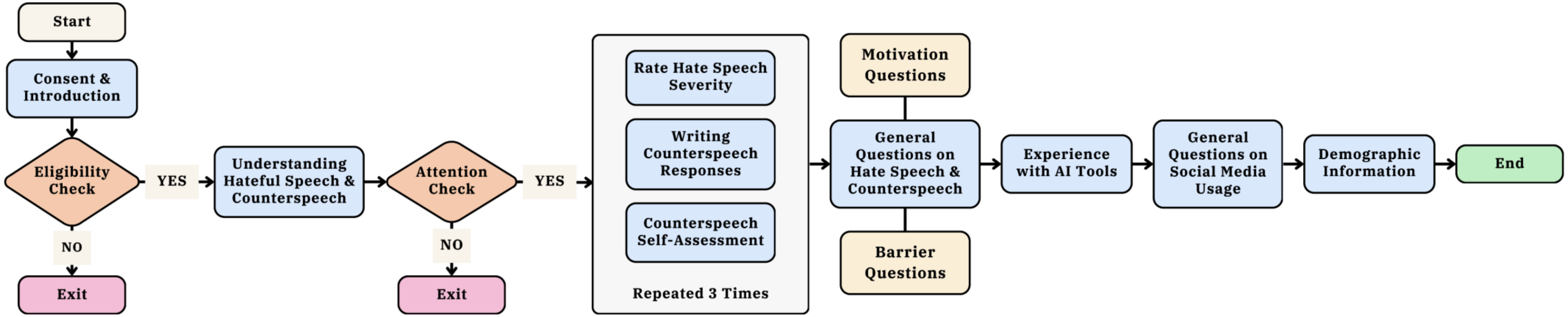
Race

Gender

Religion

Sexual Orientation

Disability



Start

Consent & Introduction

Eligibility Check

YES

NO

Exit

Understanding Hateful Speech & Counterspeech

Attention Check

YES

NO

Exit

Rate Hate Speech Severity

Writing Counterspeech Responses

Counterspeech Self-Assessment

Repeated 3 Times

Motivation Questions

General Questions on Hate Speech & Counterspeech

Barrier Questions

Experience with AI Tools

General Questions on Social Media Usage

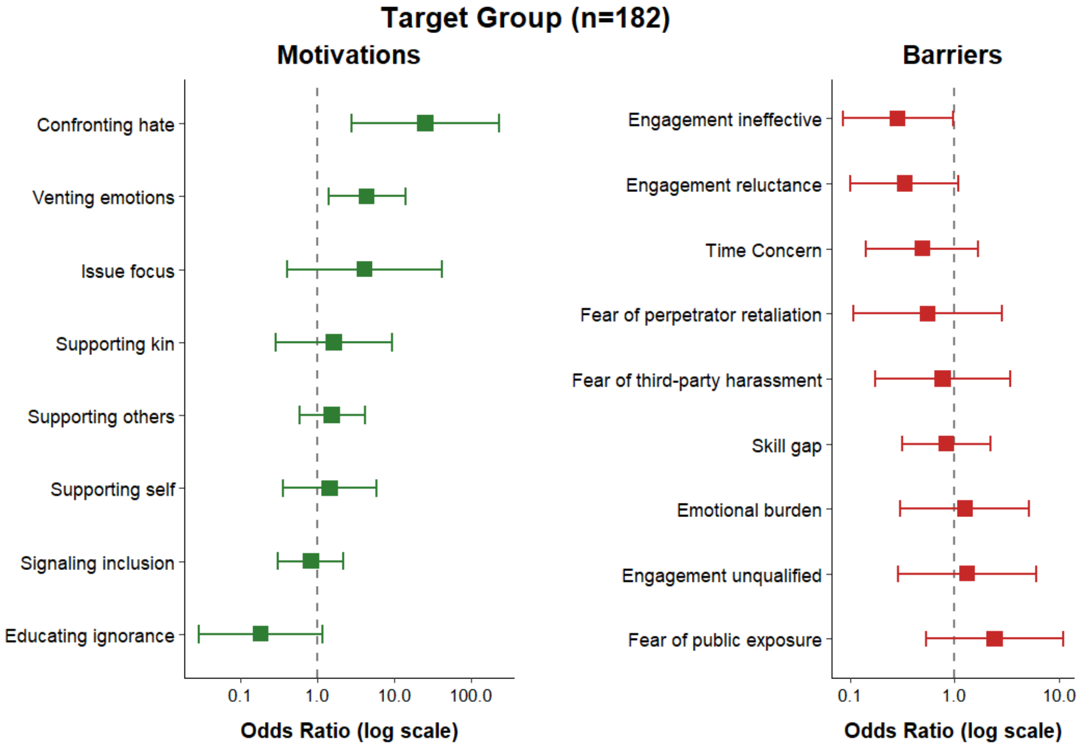
Demographic Information

End

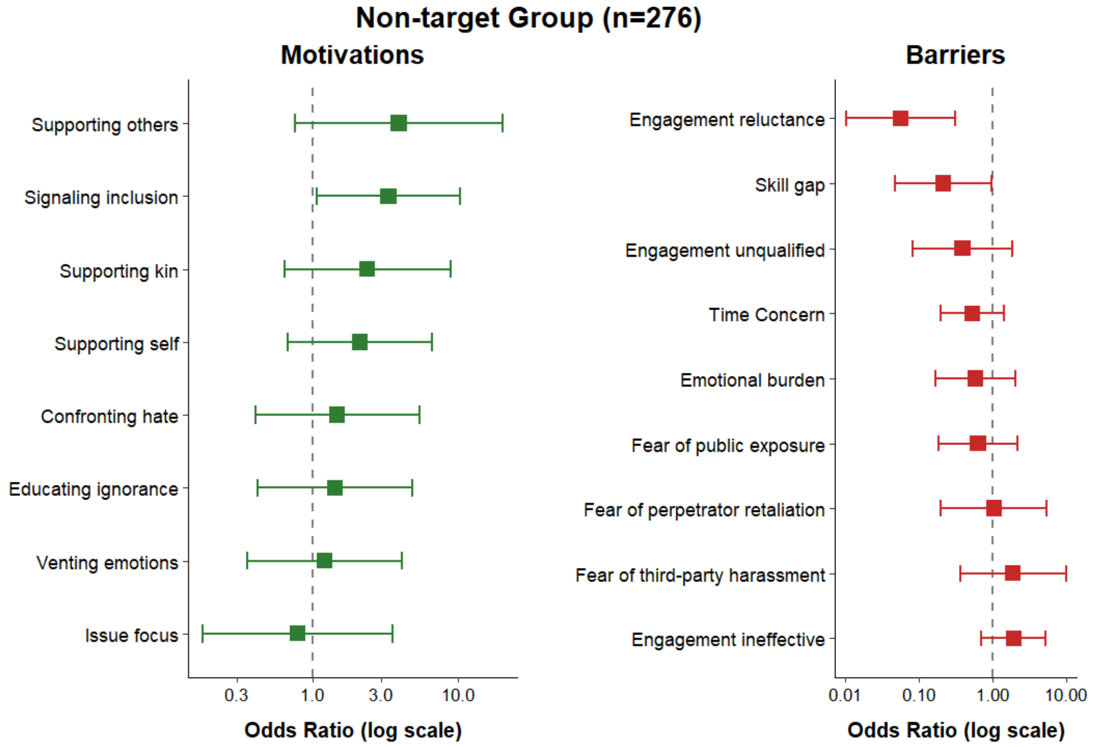
Finding 1: Past Experience is a Key Driver

Being a previous target of online hate is the strongest predictor of engaging in counterspeech. However, motivations and barriers differ significantly between groups.

If you've been a target:

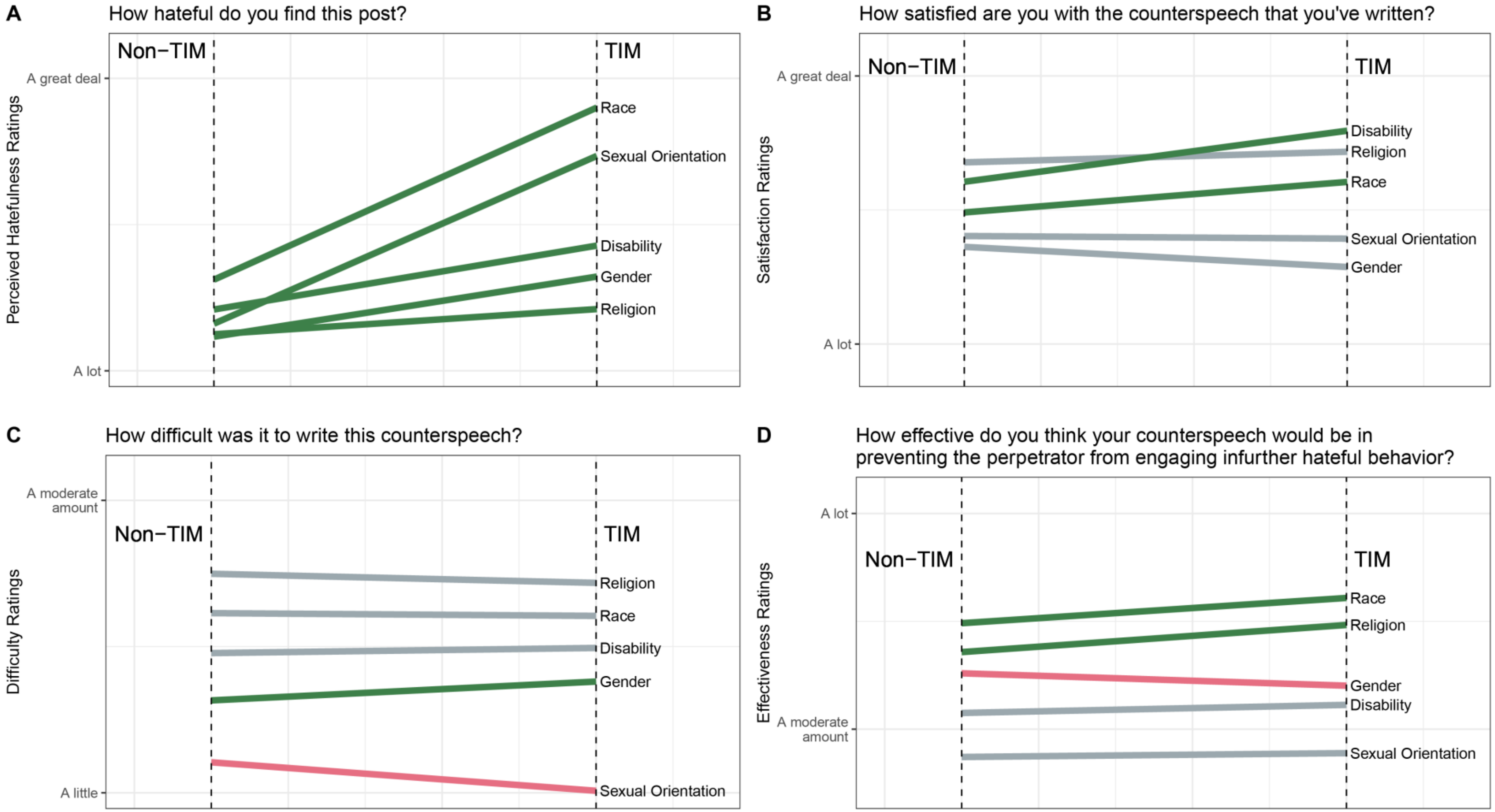


If you have NOT been a target:



Finding 2: Identity Shapes the Experience

Topic-Identity Match (TIM): When the hate speech topic aligns with the writer's identity



Finding 3: Empathy is Powerful, but Difficult

What makes a "good" counterspeech from the writer's perspective?

Length

Longer

Tone

Positive

Empathy

Understanding & Shared Feeling

The Empathy Challenge

Most effective

Most difficult to write

Conclusion & The Role for AI

Not One-Size-Fits-All

Identity and personal history are fundamental to how and why people respond to online hate.

A Clear Tension

The most effective strategies, like empathy, are also the hardest to write, creating a barrier to action.

An Opportunity for AI

AI assistants can help users craft the difficult, empathetic responses needed to effectively challenge hate online.

[1] Ping, K., Kumar, A., Ding, X., & Rho, E. H. (2024). Behind the Counter: Exploring the Motivations and Barriers of Online Counterspeech Writing. *ACM Transactions on Computer-Human Interaction*.

[2] Ping, K., Hawdon, J., & Rho, E. H. (2025). Perceiving and countering hate: The role of identity in online responses. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), 1-28.

Future Work

Our follow-up study, now under review, moves from the writer's perspective to the hateful author's, identifying the specific rhetorical strategies that successfully persuade those predisposed to hate.

Targeted Evaluation

We measured effectiveness by surveying "hate-aligned" individuals to see what actually changes their perspective and behavioral intentions.

Effective Rhetoric

Using Speech Act Theory, we found that acknowledgment and perspective-taking work, while sarcasm and accusations fail.

AI-Powered Prediction

We successfully trained language models to predict counterspeech effectiveness with over 85% accuracy, enabling scalable moderation.

Identity as Authority

The work shows why voices from targeted groups are more persuasive—they uniquely leverage lived experience through personal testimony.

Thank You!