Community Notes as Bridging Counterspeech

New Directions in Social Algorithms Research Colloquium.
October 17, 2025

James Grimmelmann
Kenny Peng

In this talk

- Why is counterspeech good?
- Why is counterspeech hard?
- What are community notes?
- How could community notes help?

Theories of free (counter)speech

The search for truth

- Speech helps society reach the truth
- So counterspeech > censorship because it is *more effective* at converging on truth
- This is a fundamentally *empirical* theory—it depends on falsifiable assumptions about how information actually spreads in society

Individual autonomy

- Speech is a core part of personhood
- So counterspeech > censorship because it:
  - … leaves *original speakers* free to speak
  - … promotes *counterspeakers'* autonomy, too
  - … lets *listeners* make up their own minds

Self-government

- Speech is a core part of democracy
- Counterspeech > censorship because it doesn't require the use of a dangerous power

Governmental vs. private speech regulations

- All of these theories agree that *governmental* speech restrictions are harmful
- But what about *private* speech restrictions?
  - Are they *restrictions* on users' speech?
  - Or are they *exercises* of platforms' speech?
- Counterspeech is a way of avoiding the issue

Counterspeech is hard

Three challenges

- Scale and speed: *economic* factors
- Reaching the audience: *social* + *technical* factors
- Persuading the audience: *psychological* factors

Scale and speed

- Misinformation has competitive advantages in the marketplace of ideas
- It is cheap and easy to produce at scale
- AI slop is not helping

Reaching the audience

- Does counterspeech reach the listeners who received the speech it responds to?
- “Falsehood flies, and the truth comes limping after it.”
- Recency and novelty
- Filter bubbles and echo chambers

Persuading the audience

- Fact checks don't seem to work
- Backfire effect
- Cultural cognition

How community notes could help

You already know about community notes

- User-provided replies to other users' content
- Algorithmic selection of replies to display
- Selects for *bridging* replies that are highly rated by diverse groups of users

Bridging speech

- *Associated Press v. United States*: "the widest possible dissemination of information from diverse and antagonistic sources is essential to the welfare of the public"
- Bridging speech does even better: it appeals to all of these "diverse and antagonistic" communities
- There are good reasons to think that it will be uniquely persuasive

A stylized example

- Consider how vaccine proponents and skeptics respond to notes about vaccine misinformation
- "Spikevax gave me horns" is unhelpful
  - It is endorsed by proponents but not skeptics
- "All reasonable scientists agree" is unpersuasive
  - It is endorsed by skeptics but not proponents
- "Donald Trump got a COVID booster" is effective because it is endorsed by both groups

Community notes as counterspeech

- Designed to *respond* to existing posts
- Designed to be *displayed* with those posts
- Platform *does not censor* underlying posts
- Platform *does not inject its own views* into community notes
- Good for truth, autonomy, and democracy

Counterspeech challenges revisited

- In theory, they're fast and scalable
- Algorithmic selection and display allows community notes to reach listeners who were exposed to the underlying speech
- Selecting for bridging makes community notes more likely to be persuasive

Final thoughts

- This is a theoretical case for community notes, not a practical demonstration that they work
- But it provides a framework for principled empirical measurement and system design
- Pay particular attention to the places in which the theories of free speech diverge

Questions