Conversational Agents for Deliberation of Harmful Content

Dhruv Agarwal, Farhana Shahid, Aditya Vashistha

Cornell University

x.com/agdhruv

e2e encrypted platforms:
>3 billion users

Group chats

WhatsApp fake news during Brazil election 'favoured Bolsonaro'

Analysis suggests vast majority of viral messages with false information were rightwing

Guardian

Viral WhatsApp Messages Are Triggering Mob Killings In India

July 18, 2018 · 9:12 AM ET

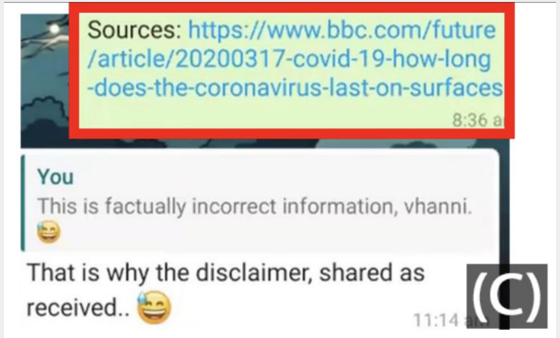
NPR

Harmful content in e2e platforms is a *grave problem*

Resende et al. WebSci 2019.
Varanasi et al. CHI 2022.

Current approaches fail to combat harmful content in this context

Admins moderate



Group members moderate

Fact-checking

- Strong in-group ties
- Deference to elders
- Non-confrontational

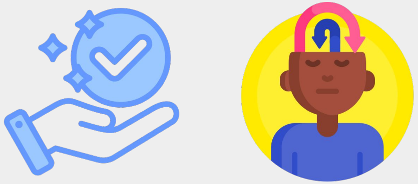
Fact-checks not circulated

Session 4d yesterday

Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines

Bago et al. 2020.

Open and inclusive discussion



Cohen 1989.

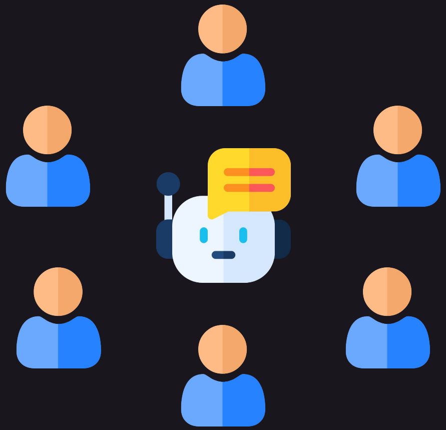
Physical deliberation spaces in rural India

Varanasi et al. CHI 2022.

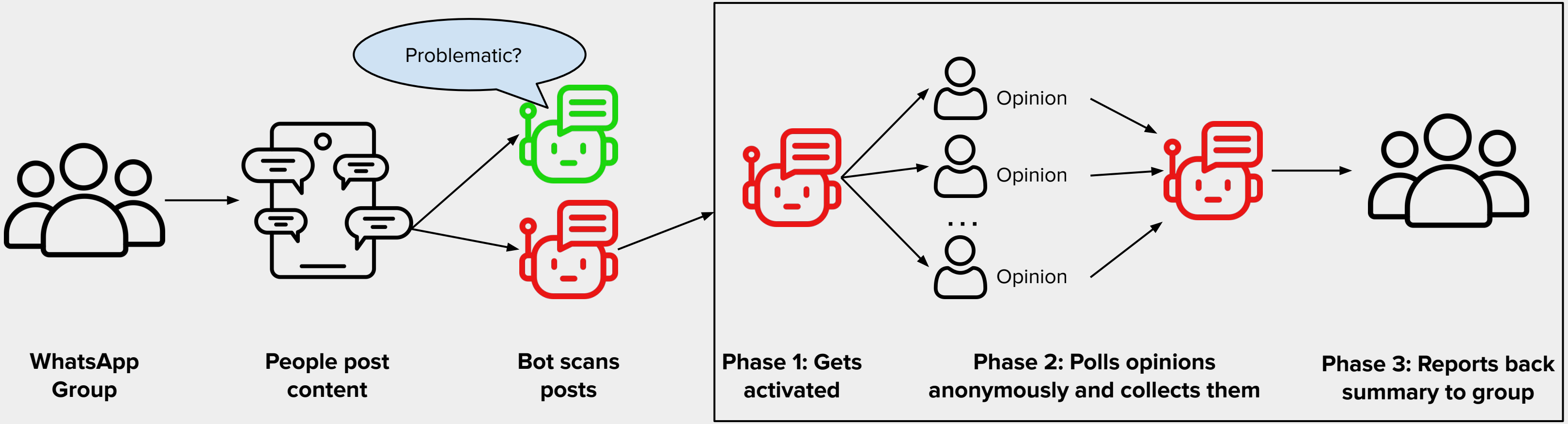
Yet no platform support for deliberation in group chats

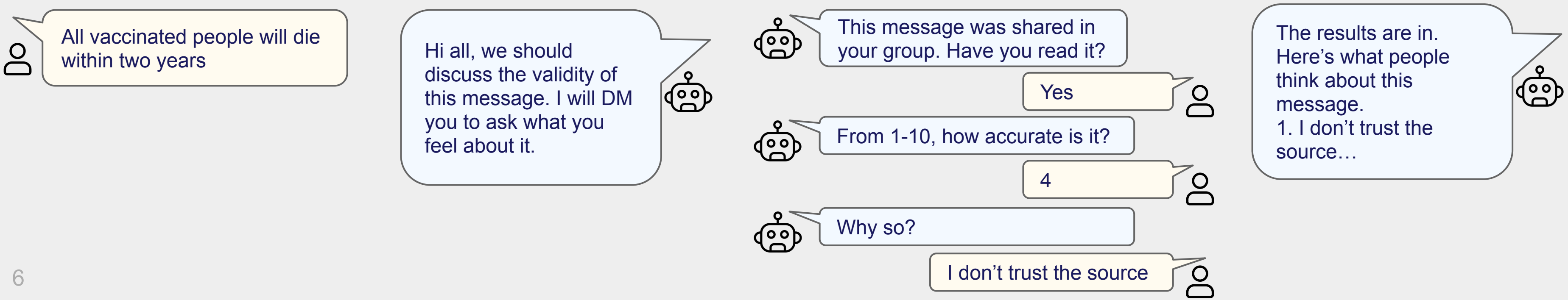
How do we design a conversational agent to facilitate *deliberation* in groups chats?

Using WhatsApp as an example platform

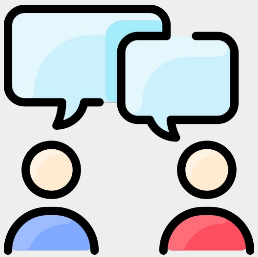


Design Probe

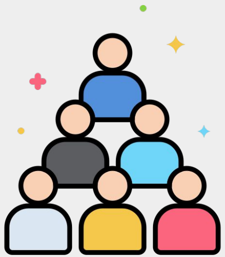




Study Details



Semi-structured Interviews



21 Participants in India

Urban + Rural
12 female
20-42 y/o

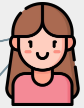
Recruitment Strategy

Snowball
Local NGOs

Findings

Anonymity important to circumvent social and cultural hierarchies

"This will allow me to speak up in my children's school group, where the School Headmaster always posts political propaganda. " – P15

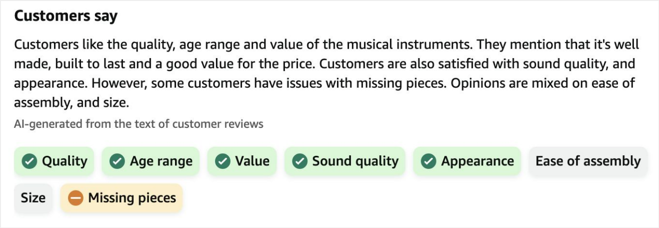


Use AI to *rephrase* opinions to avoid identification

Garble linguistic patterns, standardize grammar, remove emojis

Go further: *summarize* opinions into a "verdict"

✅ Keep well-reasoned opinions backed by evidence
❌ Filter out divisive language

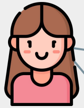


Deliberation → workload

But human-AI collaboration can help reduce the effort

Help in finding reliable information

Bias:
Which sources are “credible”?
Which content does it find problematic?



“Can you cut out the verification part, the part that my grandmother can’t do by herself.” – P10

AI writing assistance to reduce effort

Help write *constructive* arguments in an emphatic tone

Strengths and Pitfalls of Anonymous Deliberation

✅ Diversity of opinions

"The most extremist voices are the loudest, and that does not represent the group. There are enough people who believe the other way, and just knowing that fact can help." – P2

✅ A neutral mediator

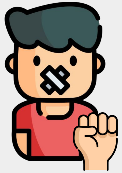
Instead of attacking the person who said it, keeps the focus on what was said

❌ May harm group dynamics, lead to in-fighting

❌ Futile for echo chambers

"Not useful for groups that relish in sharing problematic content."

Tensions in designing such an agent



vs

Should an agent call out hate speech in informal conversations?

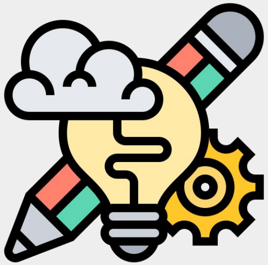
vs

Should users have to put in effort to combat harmful content online?

vs

Revealing facts vs revealing opinions

Contributions



Design space
of agents for deliberation

Efficacy of deliberation
for combating harmful content

Theoretical grounding
Deliberative theory lens

← Paper

dhruv@infosci.cornell.edu
x.com/agdhruv

More in the paper!

Design elements

Moderation,
Fact-checking,
Deliberation

Deliberative theory

Personalization: Agents as deliberative *partners*?

Pose counter-questions

Encourage critical thinking