Bystanders of Online Moderation: Examining the Effects of Witnessing Post-removal Explanations

Shagun Jhaver, Rutgers University
Himanshu Rathi, Rutgers University
Koustuv Saha, UIUC

CHI 2024
Surfing the World

Short Paper

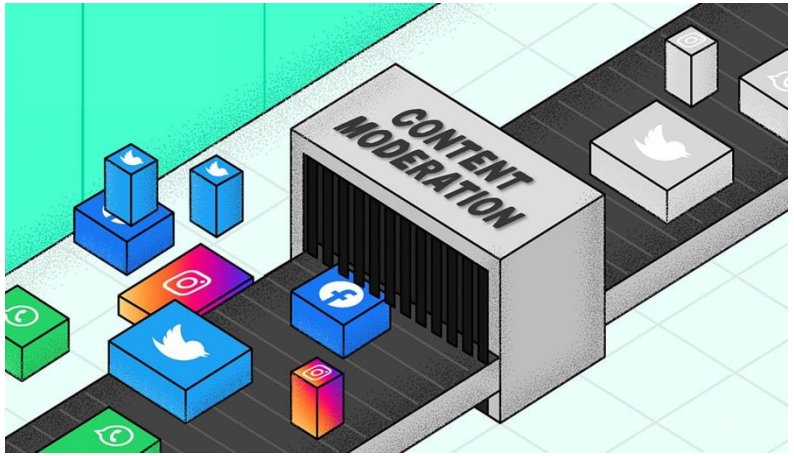
RUTGERS
School of Communication and Information

ILLINOIS

Background

Content Moderation

- Systems designed by social media platforms
- Regulate inappropriate user behaviors
- Impose measures like removing content, banning users
- AI-driven tools



Transparency in Content Moderation

- Transparency is a key guiding principle to shape social media procedures.
- We focus on transparency in end-users' experiences with moderation processes.
- Context: post removals on Reddit

Removal Explanation Example

How do Removal Explanations Affect Sanctioned Users?

Receiving Explanations => Improvements in User Attitudes[1]

1. Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman, ““Did You Suspect the Post Would be Removed?”: Understanding User Reactions to Content Removals on Reddit,” Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 192 (November 2019), 33 pages. DOI: 10.1145/3359294

How do Removal Explanations Affect Sanctioned Users?

Receiving Explanations => Improvements in User Attitudes[1]

Receiving Explanations => Improvements in User Behaviors[2]

1. Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman, ""Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit," Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 192 (November 2019), 33 pages. DOI: 10.1145/3359294
2. Shagun Jhaver, Amy Bruckman, and Eric Gilbert, "Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit," Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 150 (November 2019), 27 pages. DOI: 10.1145/3359252

Bystanders to Removal Explanations

- RQ: Do public removal explanations intended for the sanctioned users influence the posting behavior of bystanders to those explanations?

- Focusing on bystanders allows us to examine the impact of indirect experiences with punishment on users' behavior.

Methods

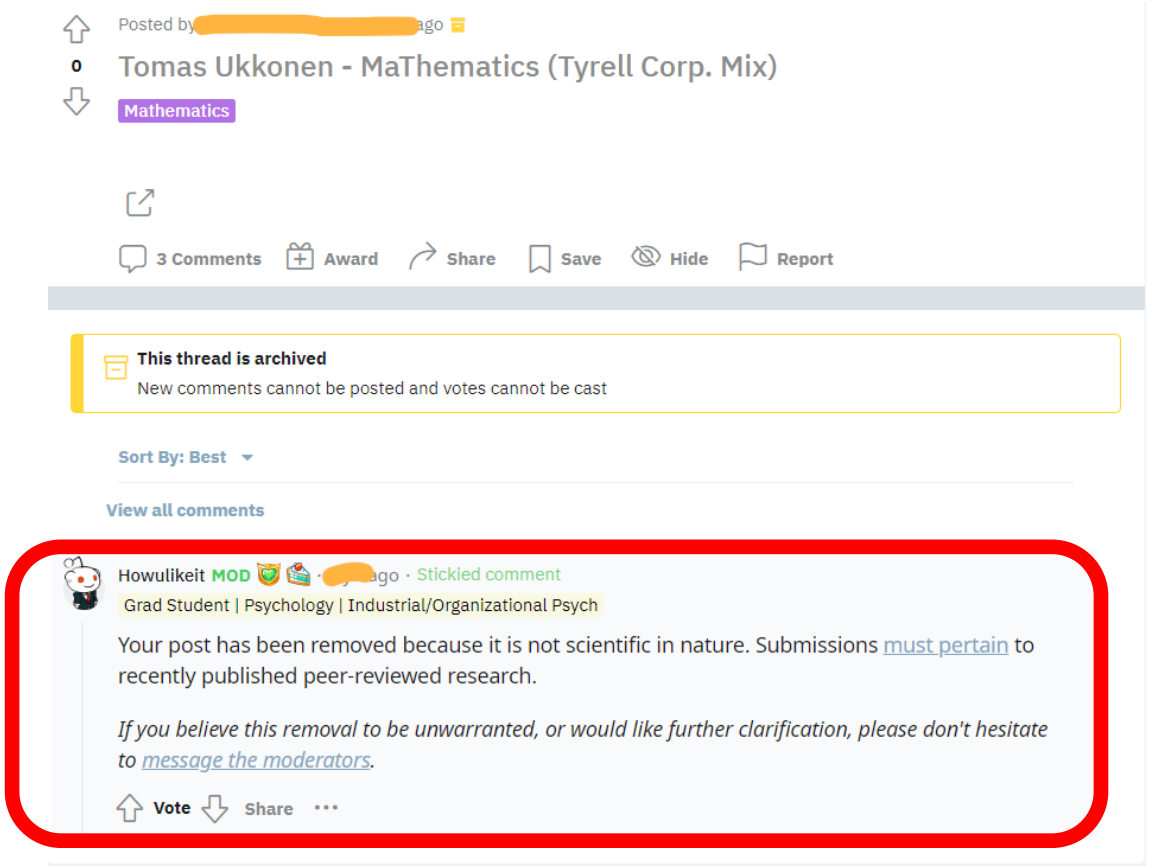
Data

- 85.5 million Reddit posts
- Two Reddit communities:
  - r/AskReddit
  - r/science
  - Largest, most active; mature moderation

Subreddit	No. Submissions	No. Comments
r/Askreddit	287,954	5,358,662
r/science	2,453	175,007

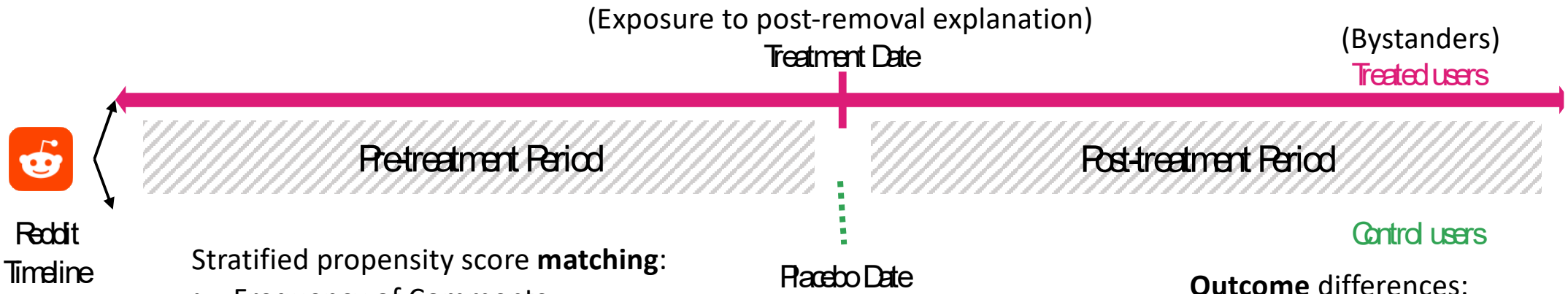
Gathering Post Removal Explanations

- Borrowed a list of 95 phrases indicating post-removal explanations from prior work[1]
- Gathered comments containing any such phrase
- Filtered for stickied moderator comments in treatment period: 1-30 June 2022
- For each thread with an explanation, all commenters become *Treatment* users
- Control users: no comment in any such thread



1. *Shagun Jhaver, Amy Bruckman, and Eric Gilbert, "Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit," Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 150 (November 2019), 27 pages. DOI: 10.1145/3359252*

Causal Inference Approach



Stratified propensity score **matching**:

- Frequency of Comments
- Frequency of Submissions
- Interactivity
- Submission Removal Rate
- Karma
- Normalized Occurrences of *n*-grams

Outcome differences:

- Posting Frequency
- Interactivity
- Post Removal Rate

Results and Implications

Results

- Bystanders (treated users) become more active after witnessing post-removal explanations.
- Bystanders increased their interactivity, i.e., they replied more to others' thread.

Outcome	ATE	Cohen's d	t-test	KS-test
r/AskReddit				
Posting Frequency	0.453	0.807	6.589***	0.640***
Interactivity	0.193	2.392	12.233***	0.960***
Post Removal Rate	0.000	0.005	0.024	0.200
r/science				
Posting Frequency	0.025	1.075	8.890***	0.515***
Interactivity	0.216	1.445	17.469***	0.879***
Post Removal Rate	0.001	0.007	0.177	0.303

Implications

- Removal explanations help boost posting frequency
  - For any removal, one moderated user but many bystanders
- Removal explanations help increase community engagement
  - Observing explanations may help learn accepted norms
- Removal explanations do not impact future post removals
  - Contrast: moderated users show reduction in future post removals
  - Possible explanation: Moderated users attend to *all* community guidelines before next postings
- Explanations must be made publicly visible
- Design space of creating explanation messages

Bystanders of Online Moderation: Examining the Effects of Witnessing Post-removal Explanations

Shagun Jhaver, Rutgers University
Himanshu Rathi, Rutgers University
Koustuv Saha, UIUC

CHI 2024
Surfing the World

Short Paper

RUTGERS
School of Communication and Information

ILLINOIS