Bans v/s Warning Labels:
Examining Bystanders' Support for Community-wide Moderation Interventions

Shagun Jhaver

RUTGERS
School of Communication and Information

Content-based Harms

Harms caused by viewing inappropriate content on social media platforms.

I focus on three content-based harms:

- *Hate speech*
- *Violent content*
- *Sexually explicit content*

Exposure can induce PTSD, depressive symptoms, & self-harm (Haas et al. 2011, Tynes et al. 2019)

Content-based Harms

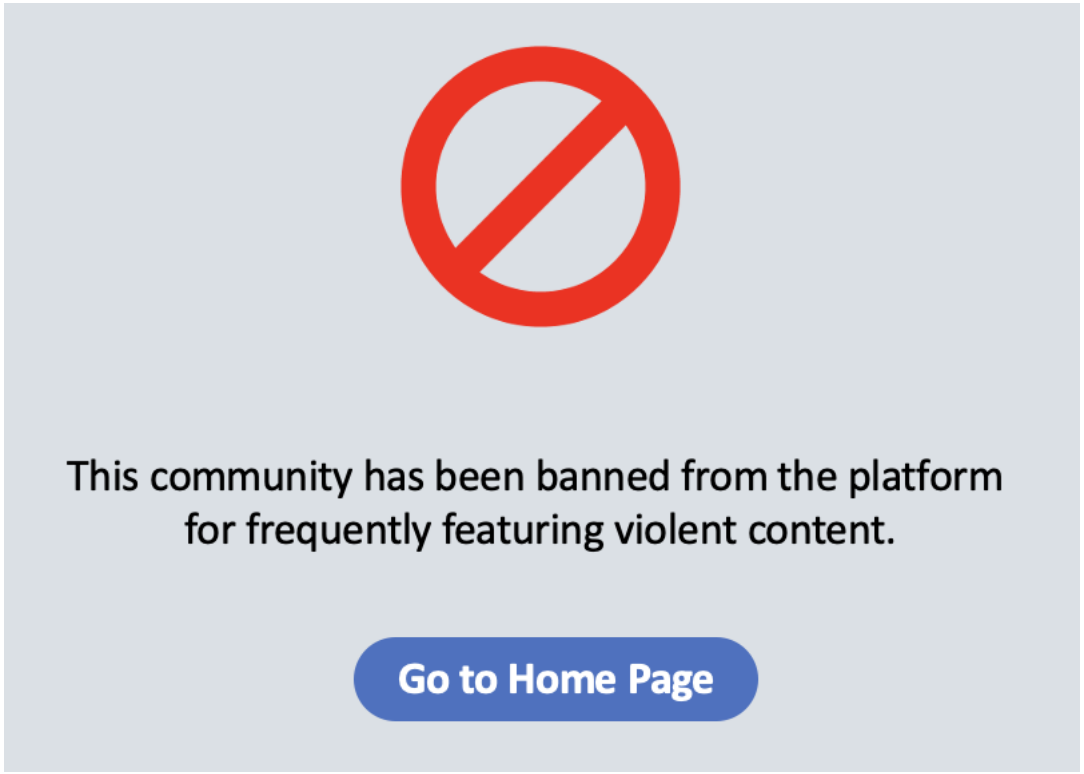
Content Moderation

User-level

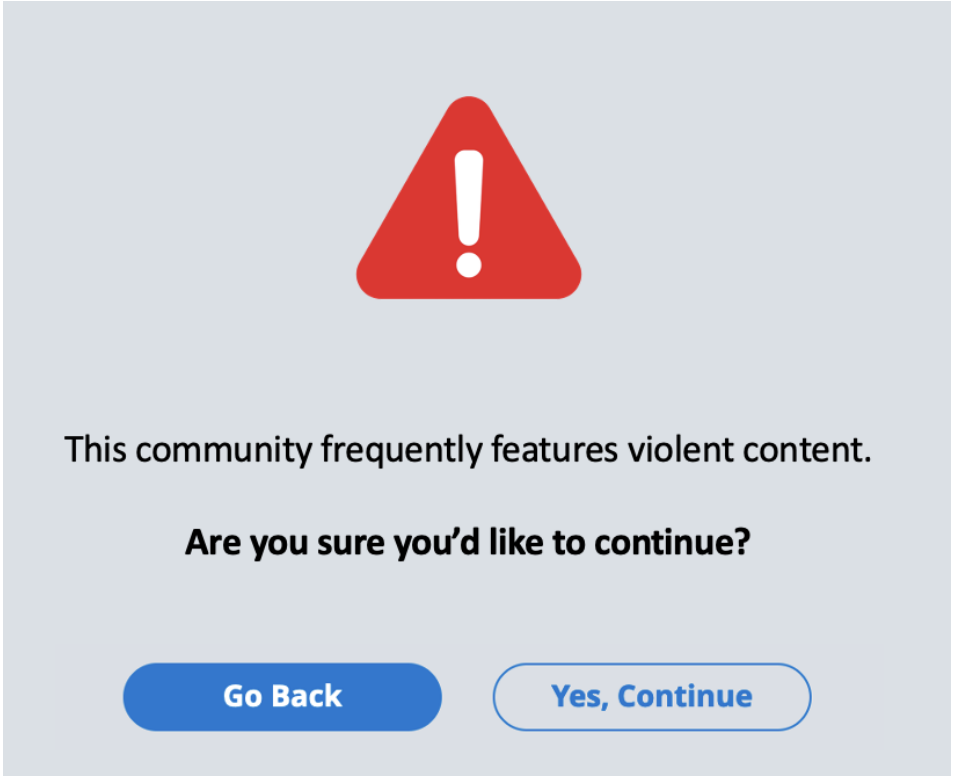
Community-wide

Two Community-wide Moderation Interventions

Bans



Warning Labels



Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit

ESHWAR CHANDRASEKHARAN, University of Illinois at Urbana-Champaign
SHAGUN JHAVER, Rutgers University
AMY BRUCKMAN, Georgia Institute of Technology
ERIC GILBERT, University of Michigan

Should social media platforms override a community's self-policing when it repeatedly break rules? What actions can they consider? In light of this debate, platforms have begun experimenting with softer alternatives to outright bans. We examine one such intervention called quarantining, that impedes direct access to and promotion of controversial communities. Specifically, we present two case studies of what happened when Reddit quarantined the influential communities r/TheRedPill (TRP) and r/The_Donald (TD). Using over 85M Reddit posts, we apply causal inference methods to examine the quarantine's effects on TRP and TD. We find that the quarantine made it more difficult to recruit new members: new user influx to TRP and TD decreased by 79.5% and 58%, respectively. Despite quarantining, existing users' misogyny and racism levels remained unaffected. We conclude by reflecting on the effectiveness of this design friction in limiting the influence of toxic communities and discuss broader implications for content moderation.

Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels

MANOEL HORTA RIBEIRO, EPFL, Switzerland
SHAGUN JHAVER, Rutgers University, USA
SAVVAS ZANNETTOU, Max Planck Institute for Informatics, Germany
JEREMY BLACKBURN, Binghamton University, USA
GIANLUCA STRINGHINI, Boston University, USA
EMILIANO DE CRISTOFARO, University College London, United Kingdom
ROBERT WEST, EPFL, Switzerland

When toxic online communities on mainstream platforms face moderation measures, such as bans, they may migrate to other platforms with laxer policies or set up their own dedicated websites. Previous work suggests that *within* mainstream platforms, community-level moderation is effective in mitigating the harm caused by the moderated communities. It is, however, unclear whether these results also hold when considering the broader Web ecosystem. Do toxic communities continue to grow in terms of their user base and activity on the new platforms? Do their members become more toxic and ideologically radicalized? In this paper, we report the results of a large-scale observational study of how problematic online communities progress following community-level moderation measures. We analyze data from r/The_Donald and r/Incels, two communities that were banned from Reddit and subsequently migrated to their own standalone websites. Our results suggest that, in both cases, moderation measures significantly decreased posting activity on the new platform, reducing the number of posts, active users, and newcomers. In spite of that, users in one of the studied communities (r/The_Donald) showed increases in signals associated with toxicity and radicalization, which justifies concerns that the reduction in activity may come at the expense of a more toxic and radical community. Overall, our results paint a nuanced portrait of the consequences of community-level moderation and can inform their design and deployment.

Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald

AMAURY TRUJILLO, Institute for Informatics and Telematics, National Research Council (IIT-CNR), Italy
STEFANO CRESCI, Institute for Informatics and Telematics, National Research Council (IIT-CNR), Italy

The subreddit `r/The_Donald` was repeatedly denounced as a toxic and misbehaving online community, reasons for which it faced a sequence of moderation interventions by Reddit administrators. It was quarantined in June 2019, restricted in February 2020, and finally banned in June 2020, but despite precursory work on the matter, the effects of this sequence of interventions are still unclear. In this work, we follow a multidimensional causal inference approach, with data containing more than 15M posts made in a time frame of 2 years, to examine the effects of such interventions inside and outside of the subreddit. We find that the interventions greatly reduced the activity of problematic users. However, the interventions also caused an increase in toxicity and led users to share more polarized and less factual news. In addition, the restriction had stronger effects than the quarantine, and core users of `r/The_Donald` suffered stronger effects than the rest of users. Overall, our results provide evidence that the interventions had mixed effects and paint a nuanced picture of the consequences of community-level moderation strategies. We conclude by reflecting on the challenges of policing online platforms and on the implications for the design and deployment of moderation interventions.

You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech

ESHWAR CHANDRASEKHARAN, Georgia Institute of Technology
UMASHANTHI PAVALANATHAN, Georgia Institute of Technology
ANIRUDH SRINIVASAN, Georgia Institute of Technology
ADAM GLYNN, Emory University
JACOB EISENSTEIN, Georgia Institute of Technology
ERIC GILBERT, University of Michigan

In 2015, Reddit closed several subreddits—foremost among them r/fatpeoplehate and r/CoonTown—due to violations of Reddit's anti-harassment policy. However, the effectiveness of banning as a moderation approach remains unclear: banning might diminish hateful behavior, or it may relocate such behavior to different parts of the site. We study the ban of r/fatpeoplehate and r/CoonTown in terms of its effect on both participating users and affected subreddits. Working from over 100M Reddit posts and comments, we generate hate speech lexicons to examine variations in hate speech usage via causal inference methods. We find that the *ban worked for Reddit*. More accounts than expected discontinued using the site; those that stayed drastically decreased their hate speech usage—by at least 80%. Though many subreddits saw an influx of r/fatpeoplehate and r/CoonTown "migrants," those subreddits saw no significant changes in hate speech usage. In other words, other subreddits did not inherit the problem. We conclude by reflecting on the apparent success of the ban, discussing implications for online moderation, Reddit and internet communities more broadly.

Research Question:

How do users perceive (1) bans and (2) warning labels for communities that frequently feature:

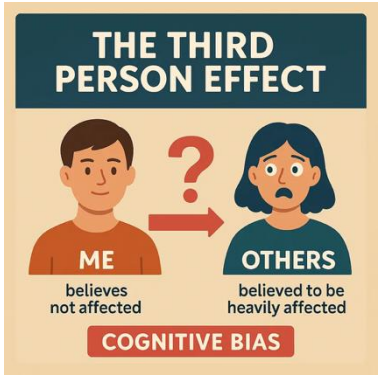
(a) Hate speech

(b) Violent content

(c) Sexually explicit content

Third-Person Effects

Support for Free Speech



Support for community-wide bans & warning labels

Methods

Survey Study

- 1,023 participants; representative sample
- Qualtrics and Lucid Theorem
- Platform-agnostic

Survey Instrument: Measures

- Third-person effects

  (1) Engaging with online communities that frequently contain hate speech would negatively influence my attitudes toward the targeted groups.
  (2) Engaging with online communities that frequently contain hate speech would negatively influence my attitudes toward anti-discrimination policies.

- Free speech support

  (a) In general, I support the First Amendment,
  (b) Freedom of expression is essential to democracy,
  (c) Democracy works best when citizens communicate in an unregulated marketplace of ideas, and
  (d) Even extreme viewpoints deserve to be voiced in society.

- Support for bans & warning labels

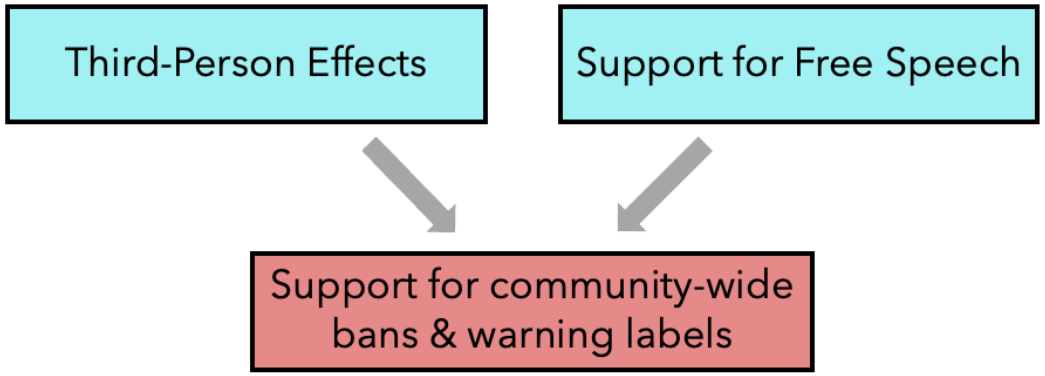
  "I support social media platforms banning any online community that frequently contains <speech category>."

Survey Questions

- Open-ended question

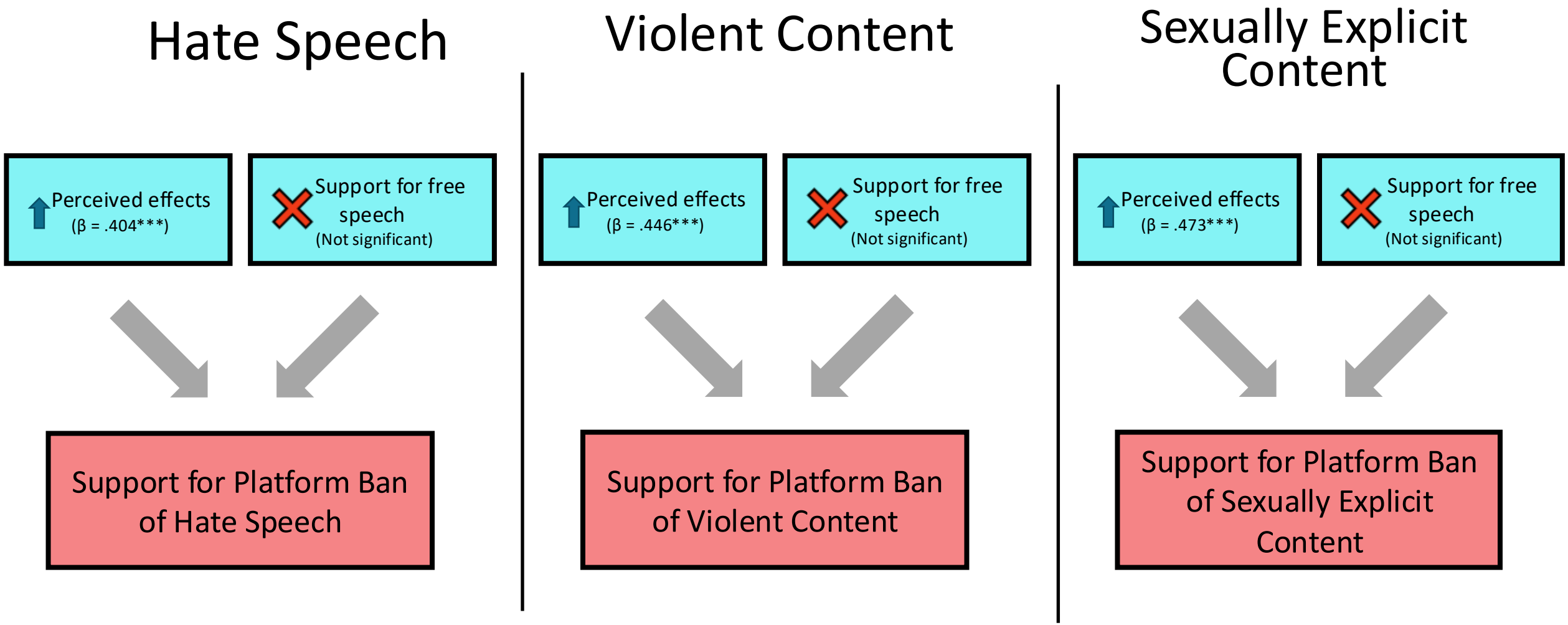
Linear Regression Models

- Independent Variables:
  - Perceived effects of speech on others
  - Support for free speech
- Dependent Variables:
  - Support for platform ban of communities with:
    1. hate speech
    2. violent content
    3. sexually explicit content
  - Support for warning labels before communities featuring
    1. hate speech
    2. violent content
    3. sexually explicit content
- Control Variables:
  - Age, gender, race, education, political affiliation, social media use

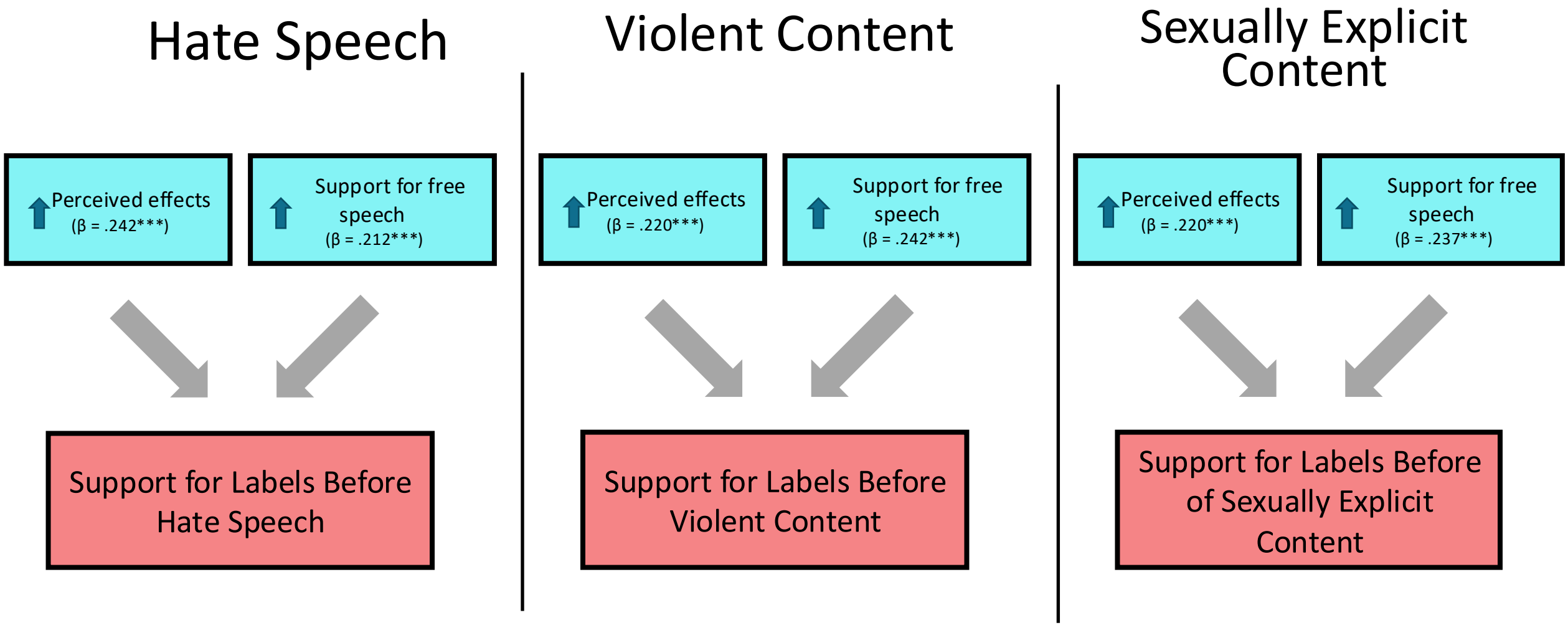


Findings

DV: Support for Community Bans



DV: Support for Warning Labels



Takeaways

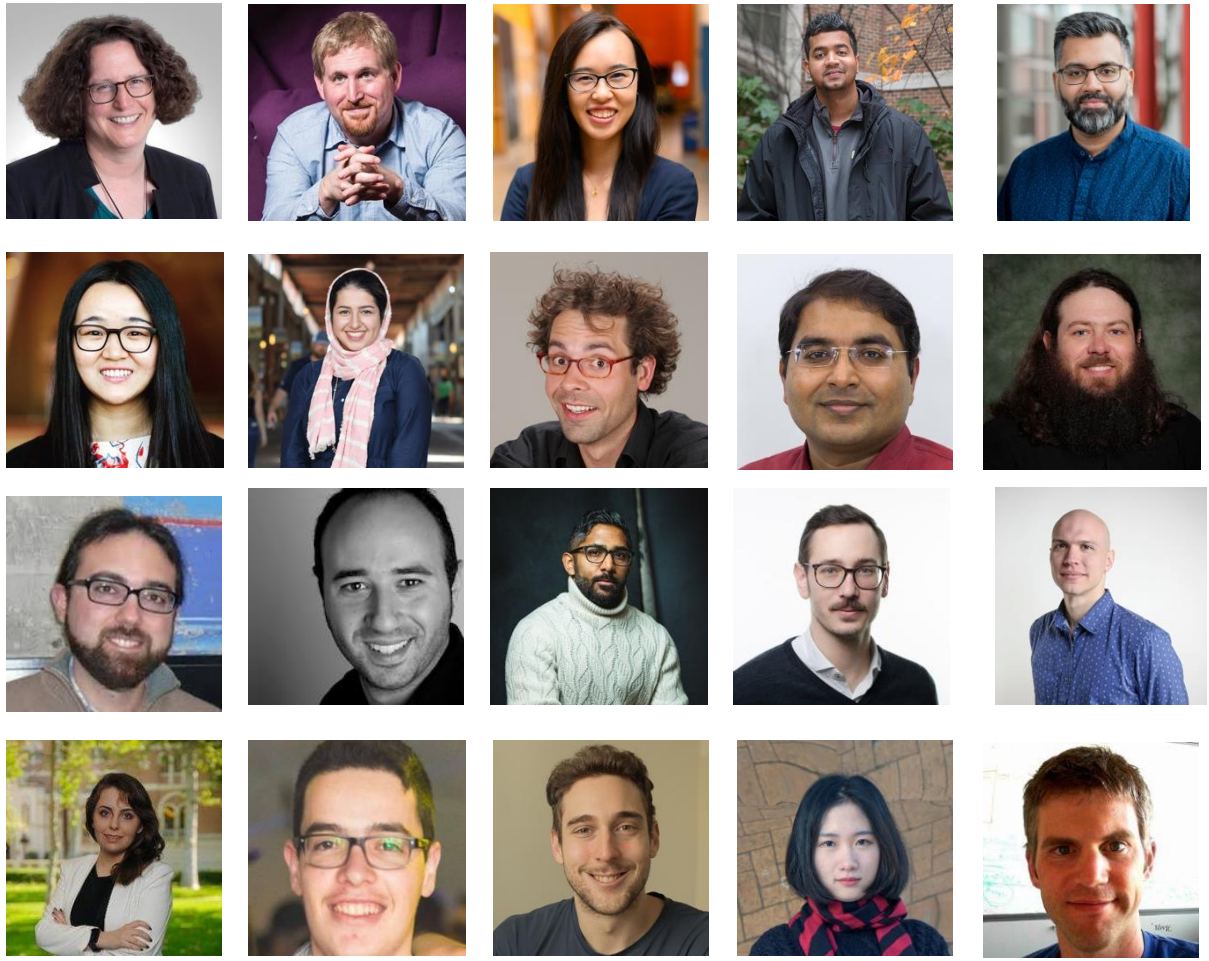
Influence of Third—Person Effects

- Perceptions of effects => support for bans and warning labels.
  - When users perceive communities as detrimental, they desire action.
- Many worry about influences on suggestible others.
- Platforms could bolster support by detailing how content poses risks to vulnerable others.

Influence of Free Speech Support

- No relation between free speech support and ban support for each category.
  - Free speech values do not make individuals more accepting of harm.
- Support for free speech predicted support for warning labels before communities.
  - Warning labels not a violation of others' free speech, but a means to empower themselves.

Acknowledgments



Georgia Institute of Technology
HARVARD UNIVERSITY
UNIVERSITY of WASHINGTON
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
UNIVERSITY OF MICHIGAN
UT DALLAS
UCL
twitch
EPFL
Microsoft Research
Berkeley UNIVERSITY OF CALIFORNIA
BROWN
airbnb
BOSTON UNIVERSITY
Inria
MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

RUTGERS
School of Communication and Information

Bans v/s Warning Labels:

Examining Bystanders' Support for Community-wide Moderation Interventions

Shagun Jhaver

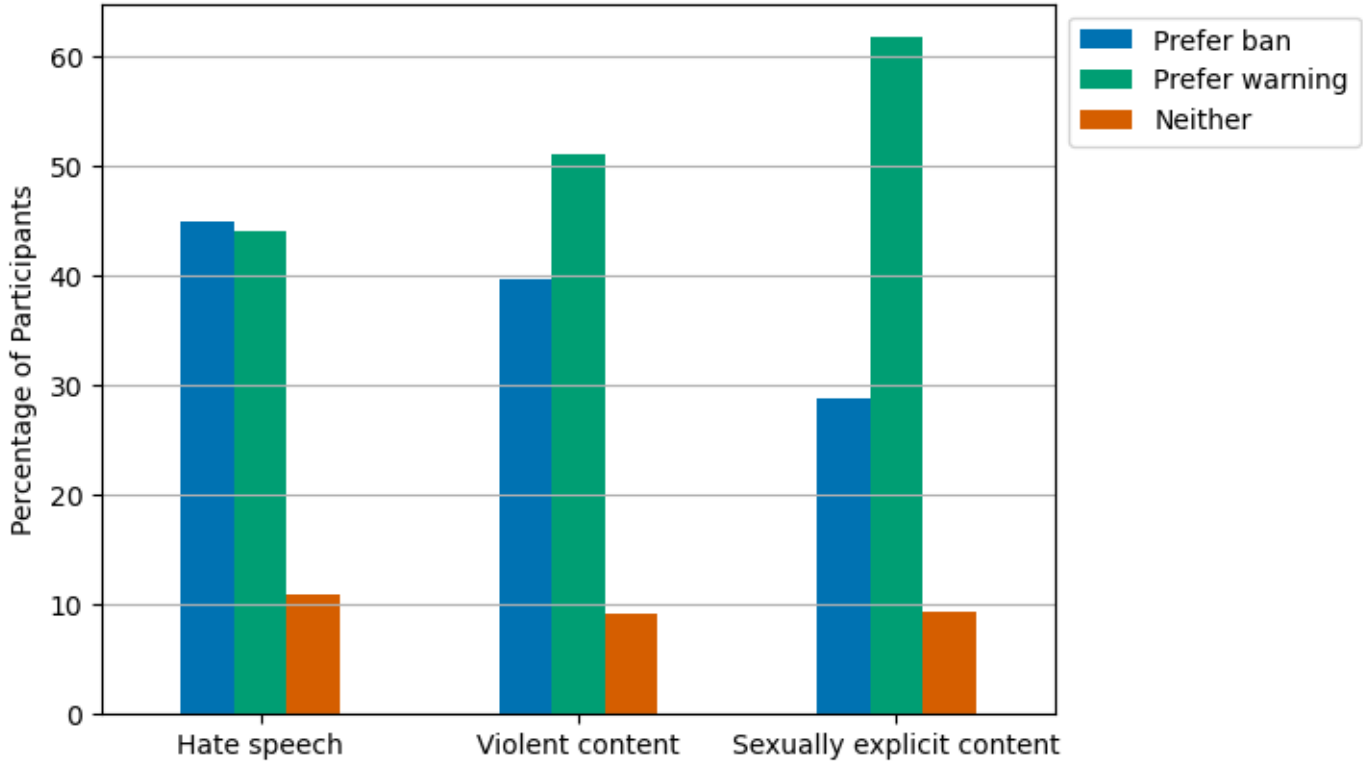
@shagunjhaver

Presented by

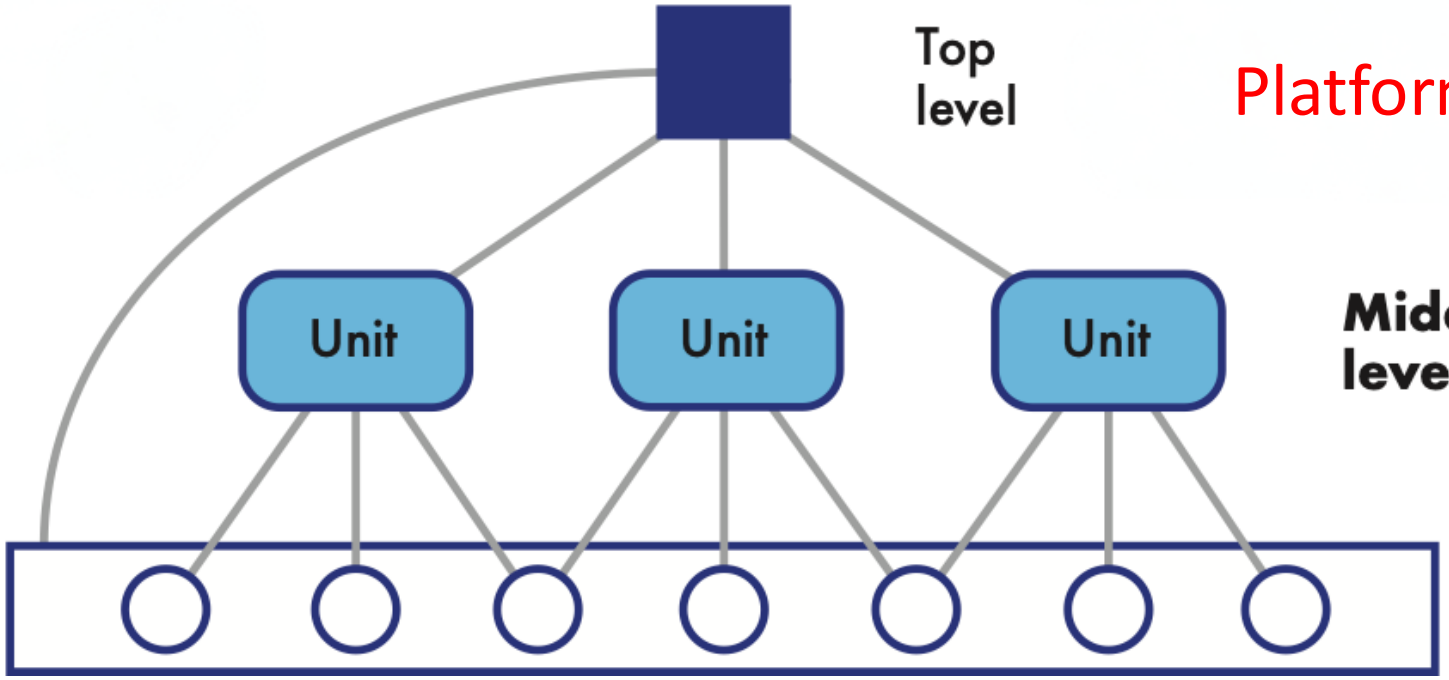
Shagun Jhaver

How should platforms handle a community that frequently feature [hate speech/violent content/sexually explicit content]?

- Preferences for bans v/s warning labels varies.
  - Users prefer moderation approaches of varying severity for different categories of norm violations.
- Only about 10% selected neither intervention.
  - This acceptance should empower community moderation, but only when warranted.



Multi-level Governance



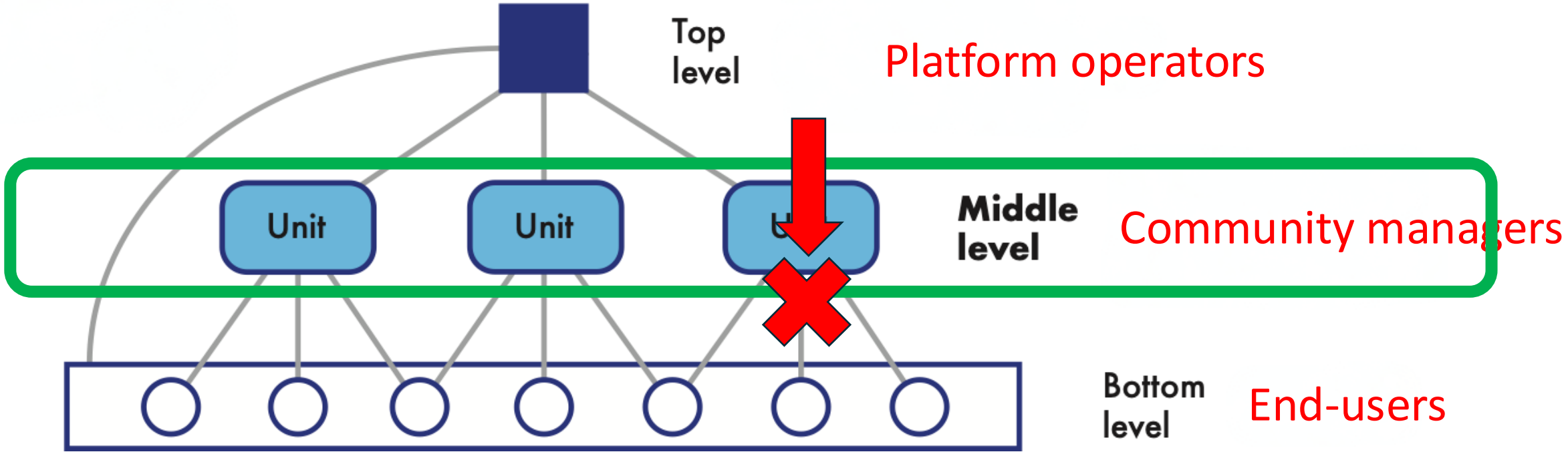
Platform operators

Community managers

End-users

Source: Shagun Jhaver, Seth Frey, and Amy X. Zhang (2023), “Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms,” Social Media + Society, 9(4).

Multi-level Governance



Source: Shagun Jhaver, Seth Frey, and Amy X. Zhang (2023), “Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms,” Social Media + Society, 9(4).