

Human-Machine Collaboration for Content Regulation

The Case of Reddit Automoderator

Shagun Jhaver, Iris Birman, Eric Gilbert, Amy Bruckman

A post on r/science, a Reddit community

 94.1k 



MMR vaccine does not cause autism, another study confirms ► [cnn.com](#)

8 months ago by [pipsdontsqueak](#)   3

2606 comments share save hide [give award](#) report crosspost

 [-] [REDACTED] 20 points 8 months ago 

 | No shit.

- Everyone

A post on r/science, a Reddit community

 94.1k 



MMR vaccine does not cause autism, another study confirms ▶ cnn.com

8 months ago by [pipsdontsqueak](#)   3

2606 comments share save hide [give award](#) report crosspost

 [-] [REDACTED] 20 points 8 months ago

 | No shit.

Removed

- Everyone

A post on r/science, a Reddit community

94.1k

MMR vaccine does not cause autism, another study confirms ► cnn.com

8 months ago by [pipsdontsqueak](#)   3

2606 comments share save hide [give award](#) report crosspost

[-] [✓] 20 points 8 months ago

 | No shit.

Removed

- Everyone

Content Moderation

A post on r/science, a Reddit community

 94.1k 



MMR vaccine does not cause autism, another study confirms ▶ cnn.com

8 months ago by [pipsdontsqueak](#)   3

2606 comments share save hide [give award](#) report crosspost

2,606 Comments!

Challenge of scale in Content Moderation

- Difficult for human moderators to review all comments
- One solution: Automated tools
- Many sites are using these tools

Do automated moderation tools work?

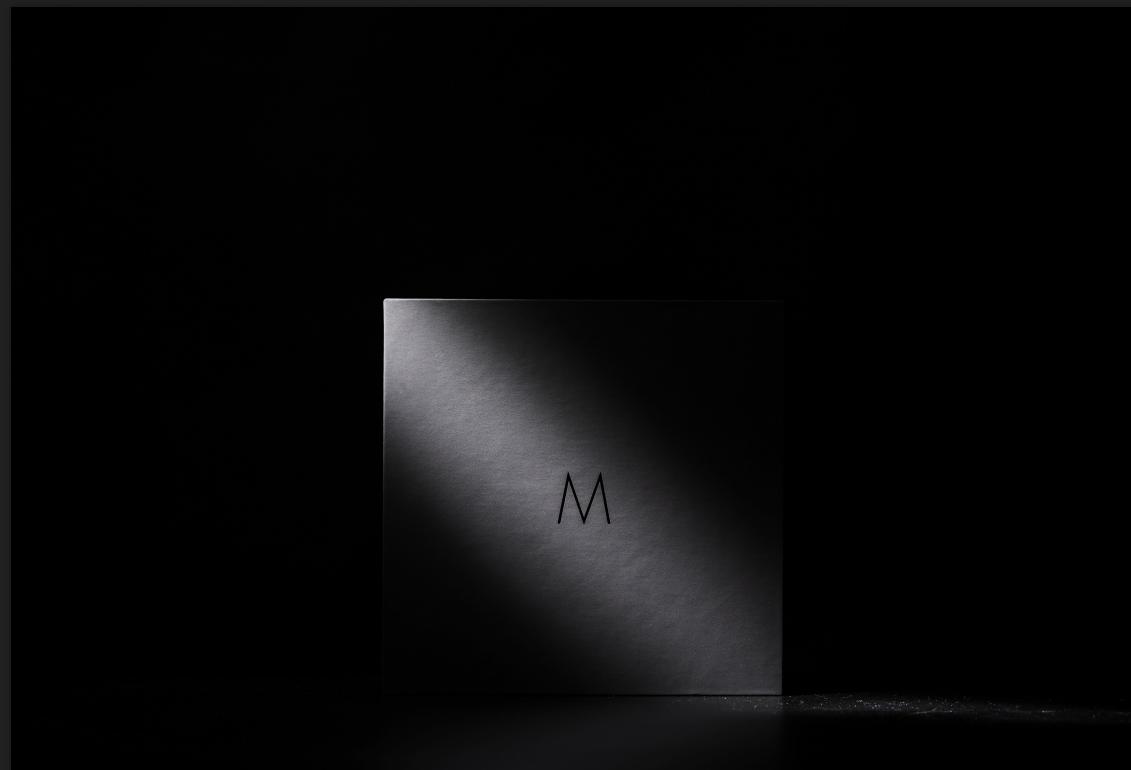
Main Contributions

- Automated tools help but they pose new challenges
- Human-machine mixed-initiative systems

Why study automated moderation?

- They perform large proportions of moderation actions
- Need to examine the sociotechnical practices of how human moderators use automated tools

Black-box content moderation systems





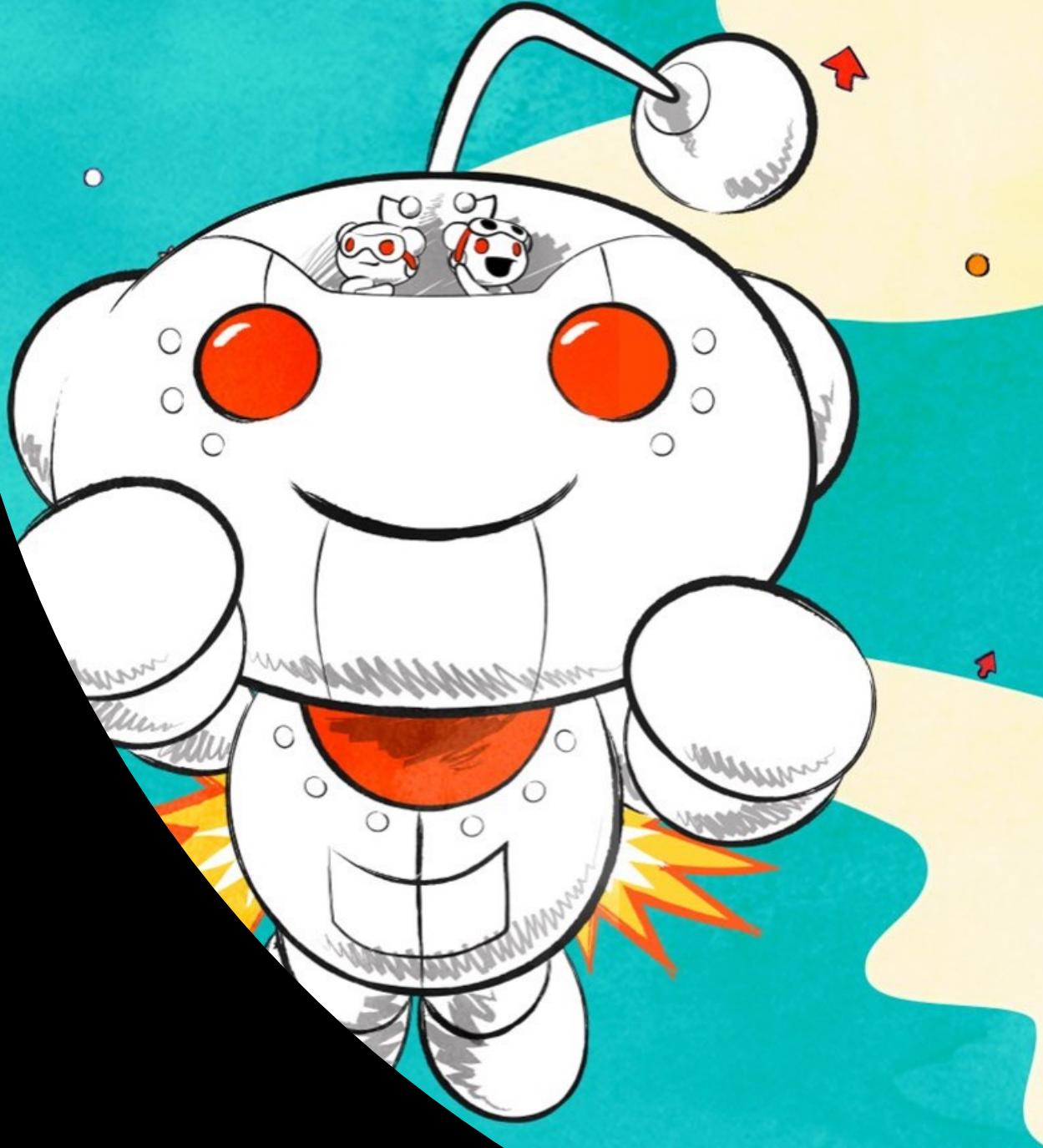
reddit

Research Questions

- How are automated tools used to help enact content moderation on Reddit?
- How does the use of automated tools affect the sociotechnical process of moderating Reddit content?
- What are the benefits and challenges of using automated tools for content moderation on Reddit?

Reddit Automod

- Most popular automated tool on Reddit
- Available to all moderators by default
- Each subreddit has its own Automod settings
- Allows moderators to configure action by setting up rules



Automod rule example

```
# Removes comments with slur/racist phrases
```

```
type: comment
```

```
body (includes): [beaner|spick|faggots|wetback|gook]
```

```
action: remove
```

Method

- Semi-structured interviews with mods of:
 - r/photoshopbattles
 - r/space
 - r/oddlysatisfying
 - r/explainlikeimfive
 - r/politics
- Chad Birch, creator of Automod
- Interpretive qualitative analysis

Findings

Automod helps reduce moderators' work

- Indispensable tool
- Executes a large amount of the menial work

“Extensive Automod rules is the only reason it’s possible to moderate ELIF.” – ELIF₁

Automod educates users

[P] [-] AutoModerator  [M] 1 point 1 year ago

Your submission has been automatically removed because titles must begin with "PsBattle: ". This needs to be exact. Please refer to the [Submission rules](#) as to why this is the case and check to see that any other rules are satisfied if you choose to resubmit.

I am a bot, and this action was performed automatically. Please [contact the moderators of this subreddit](#) if you have any questions or concerns.

[permalink](#) [embed](#) [save](#) [give gold](#)

Automod helps enforce subreddit rules

- Each subreddit has rules
- Automod is great at enforcing some subreddit rules
- Not so great at enforcing others

Automod is great at enforcing these rules:

Or/ELI5: “All Posts Must Begin With “ELI5”

Or/politics: “Do not use “BREAKING” or ALL CAPS in titles”

Automod is NOT great at enforcing these rules:

Or/ELI5: “Explain for Laypeople”

Or/ELI5: “Explanations Must Be Objective”

Same word, different contexts

“Somebody talked about how they’ve “read this shit” in an explanation of their longstanding fascination with a topic. People that use that word tend to use it in mean-spirited or unserious comments, but this was an example where it’s just for emphasis.” – Space₁

Working with Automod

- Configure Automod:
 - Automatically make decisions that are less ambiguous
 - Division of work

- “*Automod does a lot of filtering of the worst stuff for us...It makes things easier and less stressful. We don’t have to be trolling every thread for the worst stuff to get removed.*” – Space₁

- Reduces emotional labor

Technical challenge in using Automod

“It’s not necessarily user-friendly. . . it almost entirely functions on regex, and its own little quirks and syntax to implement things so it can take some time for people to get decent at using it. There are a lot of mods whose eyes glaze over when having to work with it and [they] would rather do something else.” – Pol₁

Lack of performance data

- Automod does not provide feedback on rules triggered
- Mistakes not easily caught

“A poorly phrased regex bit can make something that looks like it shouldn’t trigger on a post, trigger. But ... how do I know which one of the thirty-five Automod rules did it? How do I know which part of the post made the trigger? ... I want to know which rules were invoked for which posts, how frequently, etc. - both in aggregate and on individual posts.” – ELIF₂

Automod creates new tasks

- Regular Updating of Automod Rules
- Preventing Users from Circumventing Automod
- Correcting False Positives

Key design implications

- Facilitate development and sharing of automated mod tools

Key design implications

- Facilitate development and sharing of automated mod tools
- Build audit tools that provide visibility

Key design implications

- Facilitate development and sharing of automated mod tools
- Build audit tools that provide visibility
- Deficiencies of automated tools + Careful human administering => Improving mixed-initiative systems

Takeaways

- For platform creators:
 - A reference point for how mixed-initiative systems can be built
- For designers of automated tools:
 - Create tools that are easily understood
- For scholars of platform governance:
 - Automated tools may consistently censor certain viewpoints
- For content moderators:
 - Prepare to learn how to use automated tools

Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator

Shagun Jhaver, Iris Birman, Eric Gilbert, Amy Bruckman

Thank You!



@shagunjhaver



jhaver.shagun@gatech.edu

