Community-Centered Design of Automated Content Moderation

October 2021 - September 2024

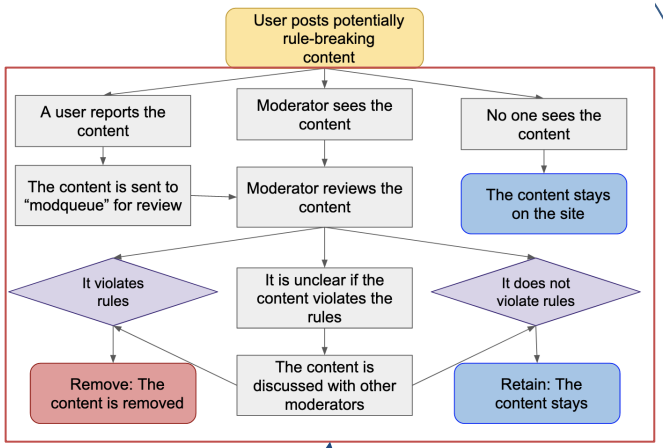
UNIVERSITY OF MARYLAND

Challenge:

Current auto-moderation tools flatten complex and nuanced moderation practices or reiterate abusive language and silencing of marginalized populations in online communities

Approach:

- Develop value-sensitive design (VSD) techniques to inform development of content moderation tools based on machine learning (ML) and natural language processing (NLP)
- Develop a learning algorithm that balances high accuracy on the moderation task and explanations that are consistent with human justification



Novel VSD-informed ML tool supports human moderation decisions

Scientific Impact:

- New design methods for ML tools that adapt to complex written policies and identify unwritten social norms
- Advance knowledge of how "machine-in-the-loop" moderation (where automated tools support moderation decisions) impacts moderator working conditions and online participant experiences

Broader Impact and Broader Participation:

- Enable better working conditions for online moderators
- Support online communities through strong policy enforcement
- Support future responses to upcoming and currently unpredictable federal content moderation legislation
- Enhance undergraduate and graduate education about software accountability methods and approaches

NSF Project # 2131508

University of Maryland, College Park
Sarah Gilbert, Katie Shilton, Hal Daumé III, Michelle Mazurek