Desk research for Automoderator

By Aishwarya Vardhana

Who is this deck for?

- Moderator Tools team
- Human Rights and Trust and Safety
- Folks interested in ethical ML practices
- Wikimedia community
- Regulators or third parties

Why does this project matter?

#1 We are in the early days of designing, building, and integrating bounded automated systems that augment the volunteer efforts of Wikipedia’s moderators. **How we do this work sets the tone** for other teams and similar efforts in the future**.**

#2 Government regulators are increasingly expecting platforms to quickly remove hate speech, misinformation, and other illegal content,. **Scaling the work of our moderators is urgent.**

What are the goals of this desk research?

1. Identify ethical considerations that should be taken into account at the very beginning of the project
2. Define principles for designing a trustworthy automated system
3. Identify mistakes we should avoid
4. Identify best practices while designing a human-on-appeal system
5. Identify interaction design challenges for moderator tools like Automoderator

Readings

- "Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator"
- AI Fairness Checklist - Microsoft Research
- Inclusive product development playbook - Wikimedia
- “Using ‘safety by design’ to address online harms”
- Data & Society — Algorithmic Accountability: A Primer
- Responsible use of artificial intelligence (AI) - Canada.ca
- ‘Social loafing’ found when working alongside robots

Findings

Findings

1. **Envision:** Creating a “system vision map” could be useful for potential fairness-related harms to stakeholder groups. Who will the system give power to? Who will it take power from? Revise the system vision to mitigate potential harms, and if this isn’t possible document why. Solicit input and concerns on system vision.
2. **Transparency is key:** Transparency and accountability increases trust. Transparency is a process.
3. **Equity and bias:** Automation can have an unanticipated impact on the platform therefore assessing risks and ethics must be an ongoing effort. Consider how automation will impact equity and determine what can be done to mitigate bias and discrimination.
4. **Work *with* moderators and help moderators work together.**
5. **Design for understanding:** Documentation isn’t just for technical admins. It can be for nontechnical people in the community, members of other product teams at the Foundation, regulators and third parties.
6. **Design for ease of use.**
7. **Design for user control.**

Based on the findings, **how might we…**

… design for transparency and accountability?
… ensure transparency is at the center of our product development process?
… assess risks and ethics as an ongoing effort?
… consider how automation will impact equity and determine what can be done to mitigate bias and discrimination?
… work *with* moderators and help moderators work together?
… design for understanding?
… design for ease of use?
… design for user control?

Creating design principles for Automoderator

Our process

We used a design principle process from the NNGroup:

1. Identify core values
2. Write, compare, and iterate (Miro)
3. Consider how these values impact users
4. Identify any common tradeoffs

Design principles for Automoderator

1. Transparency
2. Human-on-appeal
3. Community control
4. Ease of use
5. Timeliness
6. Bounded

Transparency

We value **transparency**, visibility, and accountability so that all stakeholders can easily discover, understand, and audit Automoderator should they want to. Without transparency Admins may misunderstand, and therefore accidentally misuse or refuse to use Automoderator, and users impacted by its decisions may feel censored. These situations would lead to decreased trust in and use of the automated system.

We also want to ensure that Automoderator is not invisible but rather easily visible to the editing community. We also value transparency because it makes intentional abuse or misuse discoverable. AM's actions are visible, auditable and trackable; information about how the Revert Risk models is freely available and comprehensible for AM's users.

We will continuously study data and how the Revert Risk algorithm affects Wiki projects.

Human on appeal

We value a **human-on-appeal system** because it is guaranteed that Automoderator will make mistakes. It is therefore important for there to be systems in place that allow users to provide feedback to Automoderator and appeal its decisions. Automoderator is designed to function with human involvement in its processes. Decisions made by AM should be able to be overturned by humans where necessary.

Community control

We value **community control** because we want Admins who configure Automoderator to have and feel a sense of agency over the tool and their wiki. Automoderator exists to be in service of Admins and their work. *It* works *for* them, and not the other way around. We want the community in general to feel that they have control, not just the one or two users who understand how to configure it.

Ease of use

We value **ease of use** because we do not want Admins to get frustrated when configuring Automoderator or feel alienated by its presence on their wiki. We want to ensure that anyone that interacts with Automoderator as a system is able to engage meaningfully.

Timeliness

We value **timeliness**. Whether it's reverting vandalism or providing feedback when decisions are appealed, the tool should do these actions in a timely fashion.

Bounded

We value an intentionally **bounded** system. Automoderator does not perform actions or moderator functions beyond the limits we have set out for it (i.e. not operating on other forms of actionable behaviour, non-obvious vandalism, non-moderator tasks).

What tradeoffs exist, if we consider these principles?

- There may be a tension between making Automoderator *too* visible and some community members suspicion of automation. If this tension arises, what should we prioritize? The members who want the high visibility or the benefit of obscuring Automoderator one layer deep so that it can continue running without objection?
    - **Jason:** my possibly naive take: I think we should err on the side of visibility so long as we are not adding noise to existing workflow signals; eg. notification spam, filling up existing work queues with perfunctory/meaningless steps, etc. We only want to pilot with interested communities who are open to ml-based tools. By the time we are bringing this up where it might be controversial, I'd hope we would be able to demonstrate a record of mitigating the concerns
    - **Sam**: I agree - Wikimedia communities benefit transparency above all. If anything suspicion is likely to be brought about by obscuration, not visibility.

What tradeoffs exist, if we consider these principles?

- At WMF there is often a tension between building what the community explicitly asks for versus what we, WMF product staff, think they need, from a product and engineering POV. Here again, we may come across tensions between:
  - how the community wants Automoderator to behave and how we think it should behave e.g. configuration preferences
  - the amount of documentation or information the community wants vs. documentation that accessible and digestible to nontechnical people
    - **Jason**: Our guardrails/limitations should be dictated by our principals. Different communities will have different appetites for precision vs recall and we should allow for them to land anywhere on that spectrum within the bounds of our principles. I think this is where equity as a principle comes in.

What tradeoffs exist, if we consider these principles?

- Because there are threat actors who may circumvent Automoderator if they too finely understand how it works, we may need to withhold some documentation. Who makes the decision on this? And what role do model cards play here?
    - **Jason:** I don't believe withholding documentation will be of any value for this tool. All of the criteria on which we'll operate are publicly visible, and all of our code will be open source. There are no operational secrets to hide. I think that any secret sauce in the WMF-content-moderation-slash-spam-prevention-sandwich will be in other tools.
    - **Sam:** I agree - in fact I doubt that it's possible to circumvent Automoderator by understanding it - the ML component is inherently a black box of decision making, unlike an abuse filter, for example.

What tradeoffs exist, if we consider these principles?

- False positives will be reported to volunteers who WMF cannot guarantee will respond. Even though providing a human-on-appeal system is important, we have no control over if Admins respond to false positives and help impacted users. In these instances we will not intervene and will instead prioritize the sovereignty of wikis over the needs of impacted users.
    - **Jason:** I'll go do the readings and then come back to this, but this seems icky for users, but also the only way it could work in our ecosystem. We have to stick in our lane, but is there a mechanism for the foundation to step in when a community crosses some line from messy to harmful? This is meant to be a broad question, not specific to our tool
    - **Susana**: I think this is a good possible (very probable) scenario where we have to apply some of the mitigations we talked about in the pre-mortem.What if we make checking false positive revisions part of new-comer tasks? In this case, this task would be a step up from adding images and links. Maybe call it a moderator track?
    - **Sam**: I think this is necessarily one of those 'we give this to the community and we need to trust they can handle this' situations, but as Susana points out, if we see issues with this approach there are features and levers we can investigate which stop short of stepping in ourselves

What should we *not* do? (i.e. anti-patterns)

1. Build in isolation
2. Make Automoderator a black box
3. Forgo user testing
4. Forgo documentation
5. Write documentation that is impenetrable
6. Neglect reporting workflows for humans to appeal Automoderation

Going deeper into the readings

1. "Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator"
2. “Using ‘safety by design’ to address online harms”
3. Data & Society — Algorithmic Accountability: A Primer
4. Responsible use of artificial intelligence (AI) - Canada.ca
5. ‘Social loafing’ found when working alongside robots

Insight #1: Transparency and accountability increases trust

- Applying safety by design principles requires that products are designed with safety and security defaults to the strongest option and transparency and control over recommendation and communication features. **(Safety by design)**
- Humans are accepting of error in other humans, but hold algorithms to a higher standard. **(Data & Society)**
- Who is being endowed with trust has a direct relationship with where liability for decision making should fall. **(Data & Society)**
- Determining who is the trusted decision-maker between algorithmic engineers, algorithms, and users requires careful consideration of what the algorithm claims to do and who suffers from the consequences of mistakes. **(Data & Society)**
- Understand and measure the impact of using AI by developing and sharing tools and approaches **(Canada)**

Insight #1: Transparency and accountability increases trust

- Be transparent about how and when we are using AI, starting with a clear user need and public benefit **(Canada)**
- Provide meaningful explanations about AI decision making, while also offering opportunities to review results and challenge these decisions **(Canada)**
- Be as open as we can by sharing source code, training data, and other relevant information, all while protecting personal information, system integration, and national security and defence **(Canada)**

Insight #2: Transparency is a process

- Moderators may not be able to understand the reasons behind some actions taken by automated tools.
- Moderators may have to make decisions about the levels of transparency they show in the operation of automated tools—if they are too transparent about how these tools are configured, these tools may be exploited by bad actors. **(Reddit)**
- Our findings show that there is lack of accessible data on how well the automated parts of regulation systems on Reddit work. Automod does not provide any visibility into the number of times each rule has been triggered. This is problematic because rules added to Automod sometimes have unintended effects. We found that because of the absence of easy visibility into Automod behavior, moderators often have to rely on user reports to identify the occurrence of unintended post removals. **(Reddit)**
- We caution researchers and designers that although AI moderation systems are invaluable for managing many moderation tasks and reducing workload, deploying such systems without keeping any humans in the loop may disrupt the transparency and fairness in content moderation that so many users and moderators value. This is in line with speculations made by other researchers that ML driven moderation approaches are inherently risky because they may “drive users away because of unclear or inconsistent standards for appropriate behavior. **(Reddit)**

Insight #2: Transparency is a process

- We recommend that designers build audit tools that provide moderators visibility into the history of how each Automod rule affects the moderation on the subreddit. Such visibility would allow moderators to edit Automod rules if required and control its actions more closely. Audit tools could also be enhanced to show moderators the potential consequences of creating a new rule by simulating application of that rule on already existing data in sandpit type environments. If moderators are able to visualize the type of comments that would be removed by creation of a new rule, they would be better positioned to avoid crafting broad rules that result in many false positives. **(Reddit)**
- We found that Reddit moderators show some aspects of the work of Automod to their users but not others. These decisions are important in order to retain the trust of the users while at the same time ensuring that bad actors do not game the system and bypass Automod rules. **(Reddit)**

Insight #2: Transparency is a process

- Without some level of transparency, it is difficult to know whether an algorithm does what it says it does, whether it is fair, or whether its outcomes are reliable. **(Data & Society)**
- Also, in some cases, transparency may lead to groups and individuals “gaming the system.” For example, even the minimal openness surrounding how the trending feature on Twitter surfaces topics has allowed it to be manipulated into covering certain topics by bots and coordinated groups of individuals. Therefore, different contexts may call for different levels of transparency. **(Data & Society)**
- Trust means many things in different disciplines, but one sociological perspective holds that trust is the belief that the necessary conditions for success are in place. **(Data & Society)**

Insight #3: Automation can have unanticipated impact on the platform therefore assessing risks and ethics must be an ongoing effort.

- Automated tools may affect how moderators design community guidelines. Complying with such guidelines may increase the amount of work that end-users have to perform. Therefore, using automated tools may affect not only the moderators but also the other stakeholders in content regulation systems. **(Reddit)**
- We recommend that platforms carefully reflect on the anticipated ripple effects over different stakeholders when determining which automated tools they deploy in content regulation systems. **(Reddit)**
- The use of automated tools changes the work required of moderators and their relationships with end-users in important ways. As community managers inevitably move toward adopting more automated tools for content regulation, efforts to prepare moderators for such changes will be vital. **(Reddit)**

Insight #3: Automation can have unanticipated impact on the platform therefore assessing risks and ethics must be an ongoing effort.

- Designers and moderators must recognize that the use of automated regulation systems fundamentally changes the work of moderators. For example, when subreddits use Automod, moderators' work becomes constrained to adjudicate only those postings that are not caught and removed by Automod. Moreover, it creates additional tasks that require technical expertise such as regular updating of Automod rules and preventing users from circumventing Automod. We confirm that Automod sometimes adds to the work of moderators because they have to manually approve content mistakenly removed by Automod. Therefore, when moderators incorporate new automated mechanisms in their content regulation systems, they should anticipate new tasks and prepare to execute and train for those tasks. **(Reddit)**

Insight #3: Automation can have unanticipated impact on the platform therefore assessing risks and ethics must be an ongoing effort.

- The use of more complex ML tools can disproportionately increase the workload of moderators who can work with those tools. Platforms should consider that they may lose valuable moderators when moving to systems that heavily rely on ML tools. Although hard-coding moderation criteria facilitates scalability and consistency of moderation systems, such transformation of content moderation values can end up being insensitive to the individual differences of content, for example, when distinguishing hate speech from newsworthiness. These failures to address context issues can have serious consequences, e.g., persistence of misinformation campaigns on Facebook or WhatsApp that arguably contributed to violence in Myanmar. **(Reddit)**

Insight #3: Automation can have unanticipated impact on the platform therefore assessing risks and ethics must be an ongoing effort.

- Using Automod reduces the amount of work that Reddit moderators need to do and protects them from the emotional labor of scrolling through the worst of the internet’s garbage. On the other hand, it is all too easy for moderators to configure rules that are too broad in Automod. Although such a configuration catches and removes many potentially unacceptable posts and reduces the dependency on human moderators, it results in many false positives that may alienate users. We also found that human moderators are needed to frequently update Automod rules so that Automod can account for the fluidity of culture and adaptability of violators seeking to avoid detection. **(Reddit)**
- “Our participants reported that a bulk of their moderation mail contained complaints from new users about mistakes made by Automod. This creates additional work for the moderators by requiring them to respond to user complaints about the false positives of Automod’s decisions.” **(Reddit)**

Insight #3: Automation can have unanticipated impact on the platform therefore assessing risks and ethics must be an ongoing effort.

- Working together can motivate people to perform well but it can also lead to a loss of motivation because the individual contribution is not as visible. We were interested in whether we could also find such motivational effects when the team partner is a robot. **(Social loafing)**
- When the researchers investigated the participants' error rates, they found those working with Panda were catching fewer defects after they had seen the robot had successfully flagged many errors. They said this could reflect a "looking but not seeing" effect, where people engage less once they feel a colleague or resource is reliable. **(Social loafing)**

Insight #4: Consider how automation will impact equity and determine what can be done to mitigate bias and discrimination.

- These tools may make mistakes that could have been avoided owing to their limitations of evaluating the contextual details. Our findings reveal the deficiencies of Automod in making decisions that require it to be attuned to the sensitivities in a cultural context or to the differences in linguistic cues. **(Reddit)**
- Critics have argued that the currently available AI technologies are not good at understanding the context of a given post, user, or community... they may end up resulting in many false positives, that is, posts that are not problematic to the community get removed. **(Reddit)**
- Using automated approaches to identify abusive language can result in situations where the concerns of only the dominant groups are emphasized and existing structural inequalities are exacerbated... **(Reddit)**

Insight #4: Consider how automation will impact equity and determine what can be done to mitigate bias and discrimination.

- Automated moderation tools not only exacerbate biases but they also operate simply by reacting to problems, not by dealing with their root causes. Such an approach simply hides problematic behaviors such as sexism and racism instead of interfacing with offenders in meaningful ways. This may merely push the offensive users to other platforms where their bigoted views are more welcomed. In this way, current automated tools miss out on the opportunities to examine the social and psychological factors that lead to hateful discourses. We could go well beyond simply a deployment of automated tools to change offensive or uninformed users' perspectives on socially relevant issues and help shift norms in positive ways. **(Reddit)**
- Content moderation is male-dominated therefore by building this tool for moderators, we are inadvertently building mostly for men. **(Reddit)**
- Algorithms are often deployed with the goal of correcting a source of bias in decisions made by humans. However, many algorithmic systems either codify existing sources of bias or introduce new ones. Additionally, bias can exist in multiple places within one algorithm. **(Data & Society)**
- Unless algorithms are consistently monitored and adjusted as time passes, they reinforce the values they were created with and can become rapidly outdated. **(Data & Society)**

Insight #5: Work *with* moderators and help moderators work together

- We found that moderators self-assess their skills at configuring Automod, practice care when editing Automod rules, and coordinate with other moderators through external channels like Slack to resolve disagreements. **(Reddit)**
- Moderators could be asked to nominate additional features that may be informative in improving performance in those spaces. **(Reddit)**
- Platforms may also promote sharing of such tools on a centralized repository so that other moderators can directly access them and adapt them for their own communities. **(Reddit)**
- There is no central repository of all the automated tools that moderators can directly use. Moderators only come to know about such tools through their contacts with moderators in other subreddits. This results in duplicate effort on the part of bot developers. To avoid such duplication, platforms like Reddit should encourage volunteer developers to build tools that can be quickly adapted to enact regulation in other similar settings. Platforms may also promote sharing of such tools on a centralized repository so that other moderators can directly access them and adapt them for their own communities. **(Reddit)**

Insight #6: Design for understanding

- Automod not only reduces the time-consuming work and emotional labor required of human moderators by removing large volumes of inappropriate content, it also serves an educational role for end-users by providing explanations for content removals. **(Reddit)**

Insight #7: Design for ease of use

- Although Reddit provides these moderation tools to all moderators by default, many moderators find these tools inefficient and cumbersome to use. **(Reddit)**
- Moderators want the ability to quickly locate the settings that result in undesirable regulation decisions and fix them. Therefore, automated systems should be designed so that moderators have detailed visibility into how automation affects content curation. **(Reddit)**
- We also found that only a few technically adept moderators can configure Automod, and many subreddits are unable to tap into the full potential of Automod. This is similar to Geiger and Ribes' finding on automated regulation in Wikipedia that while many "workarounds are possible, they require a greater effort and a certain technical savvy on the part of their users". **(Reddit)**

Insight #8: Design for ease of use

- We recommend that automated systems be designed in such a way that moderators can easily understand and configure their settings. This would allow more moderators to engage with automated systems, and facilitate conditions where a larger share of moderators can influence content curation using automated tools. We found that only a small number of moderators in each subreddit configure Automod because others do not have the technical expertise to make such configurations. **(Reddit)**
- It is important to explore how the use of automated tools shapes the explainability of moderation decisions and the perceptions of affected users. **(Reddit)**

Insight #8: Design for user control

- Centralized moderation tools and mechanisms are often developed using universalist design principles and practices that assume that the “default” imagined users belong to the dominant social groups... instead, mechanisms that allow these moderators to develop and deploy regulation tools that meet the unique requirements of their communities can substantially improve content regulation... **(Reddit)**
- Moderators should also be able to tune the configurations of such systems at a granular level and maintain control over how these systems work. **(Reddit)**
- Our findings show that moderators adopt Automod because they can directly control how it works by editing its configuration. They can understand the mistakes made by Automod by observing the keywords that triggered those mistakes and explain such mistakes to placate dissatisfied users. Research has also shown how retaining control over content regulation is important to the moderators. **(Reddit)**
- Since automated tools are likely to perform worse than humans on difficult cases where understanding the nuances and context is crucial, perhaps the most significant consideration is determining when automated tools should remove potentially unacceptable material by themselves and when they should flag it to be reviewed by human moderators. **(Reddit)**

Insight #9: Design for user control

- We echo calls by previous studies for building systems that ensure that the interactions between automation and human activities foster robust communities that function well at scale. **(Reddit)**
- Moderators value Automod because it provides them a great level of control and understanding of the actions taken by Automod. Our findings reveal that moderators who do not understand how automated tools work may not be able to contribute as much after these tools are adopted. This can, in turn, affect the dynamics of relationships among the moderator team. This highlights the significance of creating tools whose configurations are easily understood by the moderators, and designing tutorials that assist this understanding. **(Reddit)**

Insight #9: Design for user control

- "I think, putting it in control of people directly made a big difference. . . they feel a lot better being able to know, 'Okay, I have this configuration. If it starts doing things that I don't want it to, I can just wipe this page out and it'll stop.' "—Chad, highlighting the importance of giving moderators clear control over turning Automod on/off. **(Reddit)**
- "There was also a lot of users that were quite upset about it simply because they call it basically the censorship bot because it can just remove anything immediately with no ability for people to reason with it or convince them that it's the wrong decision."—Chad **(Reddit)**
- We recommend that more platforms should consider providing API access that volunteer developers can use to build and deploy automated regulation bots that meet the specific needs of their communities. **(Reddit)**

Methods

- The Reddit research included 16 in depth, semi-structured interviews and qualitative analysis
- **How might we use the same methodology to measure the quality of our tool?**

Backlog

Ideas on how Automoderator could help Growth

- Two approaches to growth
  - Teaching people how to edit
  - Making editing easier → Structured tasks
- “Some of the resulting discussions are high-level, while others are extremely specific to individual wikis. “They’re involved in helping even designing the different algorithms for the different languages,” says principle UX designer Rita Ho — Vietnamese-language Wikipedia, for instance, needed its algorithm tweaked to account for how the language defines the beginnings and endings of words. An individual wiki’s administrators can also opt to turn the features off — although, so far, Ho and Miller say that’s been rare.”
- “the goal is to help build up the number of people who feel comfortable connecting with other humans in Wikipedia’s community, particularly in smaller wikis that badly need new editors. Systems like structured tasks are supposed to let people dip their toes in the water — but eventually, they’ll have to jump in.”
- “There are community members who are concerned that the more newcomers interact with automated processes, the less they understand the fundamentals of the wiki process, the community-based process,” acknowledges Miller. “Because these communities, even though they need images and they need links, they also need their future administrators, their future people that discuss policy, the future people that write full articles from whole cloth. And so part of our design is — how can the user realize that they want to discover more and get deeper into this?

What do Reddit moderators do?

1. Coordinating with one another to determine policies and policy changes that guide moderation decisions
2. Checking submissions, threads, and content flagged by users for rule violations
3. Replying to user inquiries and complaints
4. Recruiting new moderators
5. Inviting high-profile individuals to conduct AMA (Ask Me Anything) sessions
6. Creating bots or editing Automod rules (described below in this section) to help automate moderation tasks
7. Improving the design of the subreddit using CSS tools.

Other thoughts

- It is clear to me how the IRS is closely connected to Moderator Tools. They can potentially be part of one system of moderation. We might, at some point, ask community members to flag content that violates policies.
- Should we do take this Algorithmic Impact Assessment - Évaluation de l'incidence algorithmique?