Cura: Curation at Social Media Scale

WANRONG HE, Tsinghua University, China
MITCHELL L. GORDON, Stanford University, United States
LINDSAY POPOWSKI, Stanford University, United States
MICHAEL S. BERNSTEIN, Stanford University, United States

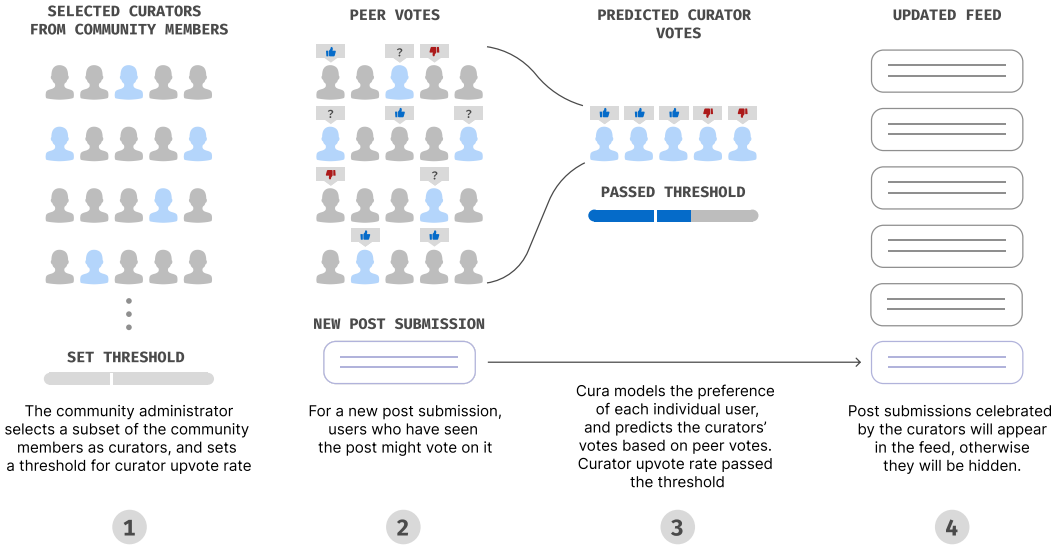

Fig. 1. Cura empowers communities to create feeds that reflect the taste and opinions of designated *curators*, from specific individuals to a democratic polity. (1) The creator selects members as curators whose upvote behavior should determine the content that should be included in the community. (2) Given other community members who have voted on a post, (3) Cura estimates the curators' preferences for the post submission. If the proportion of curators who upvoted or are predicted to upvote passes a threshold, (4) the post appears in the group feed; otherwise, it remains backstage.

How can online communities execute a focused vision for their space? Curation offers one approach, where community leaders manually select content to share with the community. Curation enables leaders to shape a space that matches their taste, norms, and values, but the practice is often intractable at social media scale: curators cannot realistically sift through hundreds or thousands of submissions daily. In this paper, we contribute algorithmic and interface foundations enabling curation at scale, and manifest these foundations in a system called *Cura*. Our approach draws on the observation that, while curators' attention is limited, other community members' upvotes are plentiful and informative of curators' likely opinions. We thus contribute a

Authors' addresses: Wanrong He, Tsinghua University, Beijing, China, hewanrong8@gmail.com; Mitchell L. Gordon, Stanford University, Stanford, United States, mgord@cs.stanford.edu; Lindsay Popowski, Stanford University, Stanford, United States, popowski@stanford.edu; Michael S. Bernstein, Stanford University, Stanford, United States, msb@cs.stanford.edu.

2573-0142/2023/10-ART337 $15.00
https://doi.org/10.1145/3610186

transformer-based curation model that predicts whether each curator will upvote a post based on previous community upvotes. Cura applies this curation model to create a feed of content that it predicts the curator would want in the community. Evaluations demonstrate that the curation model accurately estimates opinions of diverse curators, that changing curators for a community results in clearly recognizable shifts in the community's content, and that, consequently, curation can reduce anti-social behavior by half without extra moderation effort. By sampling different types of curators, Cura lowers the threshold to genres of curated social media ranging from editorial groups to stakeholder roundtables to democracies.[1]

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: social media, curation, feed algorithms

ACM Reference Format:
Wanrong He, Mitchell L. Gordon, Lindsay Popowski, and Michael S. Bernstein. 2023. Cura: Curation at Social Media Scale. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 337 (October 2023), 33 pages. https://doi.org/10.1145/3610186

1 INTRODUCTION

Curation enables us to execute focused visions for social spaces. Curated social spaces arise both online and offline: editorial boards for The Atlantic and the New York Times curate which submitted op-eds to publish, administrators at popular social news sites such as Slashdot manually select a small number of submitted tech news stories per day to publish [45], museum curators decide on which pieces of art to showcase and arrange into shows, teachers pick examples of student work to share with the class, and online publishers decide which news or comments to highlight [21, 78]. As seen in these examples and others, a curation metaphor empowers community leaders to select which content is shared with the community. By giving curators control over what community members see, curation empowers curators to strongly influence the group's descriptive norms [17, 18, 44, 58].

Curation, however, has remained intractable in many large-scale online communities. How might a community leader or moderator [67] execute a focused vision in a community that may have thousands of members or more, as on Reddit, Facebook Groups, large Slack workspaces or e-mail lists, or forums? Manually reviewing and approving every submitted post can require incredible effort: curating the NYT Picks comments in the New York Times online requires an entire dedicated, paid team [21, 78]. Already, moderators' bandwidth on these platforms is limited to reviewing only a very small proportion of content [56], and moderators are often already overworked [30]. Lacking tools to enable curation in large-scale communities, many platforms today adopt the metaphor of broadcast in which all content and replies are posted into a large public square instantaneously, then rely on the community itself to vote on or otherwise review the submissions. This broadcast metaphor has been successful because it lowers the effort threshold [53, 63] to sharing and consumption, allowing communities to gain and keep attention at scale [15]. Unfortunately, broadcast also facilitates off-topic and aggressive posts that attract attention [37] and erode pro-social norms [13, 17, 40], unraveling the goals that a curated community might seek.

To expand the availability of curation to large-scale online communities, in this paper we introduce a system called *Cura*. Cura's goal is to facilitate online communities whose content reflects the curators' opinion on what to share, and can do so at scale without curator review on every submission. Our approach is motivated by the observation that *community members'* opinions are plentiful and can provide substantial information about *curators'* opinions. By observing which community members upvote a post, we can learn to predict whether the curators will upvote the

[1]The code is available at https://github.com/StanfordHCI/Curation-Modeling.

post for curation into the community. This approach allows communities to run at scale on user contributions, while maintaining a north star of their curators' opinions. We operationalize this approach by contributing a deep learning transformer architecture, which we call a *curation model*, that takes as input the target curator(s), the community, the post, and any upvotes on the post from community members, then estimates whether the curator will upvote the post.[2] Learning patterns between community votes and curator votes, the curation model updates a post's status as it observes new upvotes from community members. We additionally contribute interactive models in Cura: a management interface for curated communities, where the community's administrators or moderators can select curators whose taste the system should model; a frontstage for posts that clear the model's curation threshold; and a backstage for the others where they are less visible but still browsable to the community for further feedback.

We evaluate Cura on existing Reddit data to test whether it accurately estimates curators' opinions, empowering the creation of different spaces by selecting different curators. We sample curators from a variety of identity and political affiliations via a large public dataset of upvotes on Reddit. We find that a curation model trained on this Reddit dataset estimates curators' votes with accuracy 81.96%, whereas the traditional majority vote (Reddit 'score') performs at 65.9%. We also observe that the curation model also estimates present-but-quieter users—lurkers as curators—accurately, enabling curation practices that more reflect the broader community and not just the most active members. We then demonstrate that applying curation to large and popular subreddits can execute dramatic shifts, for example refocusing a large technology community around human-centered concerns, or removing anti-LGBTQ+ content from a large teenagers community. In a field experiment, we confirm that these shifts are clear and recognizable to people: participants can accurately identify which versions of a community were curated in our system by specific subgroups. Finally, we measure that curation roughly halves the number of anti-social violations in a large community—without requiring any additional moderation effort.

Curation facilitates the creation of social media community types at scale that would be otherwise difficult. The contents of today's social media are influenced most heavily by the upvote behaviors of the most active users. But, by selecting different curators, communities can embed stronger opinions about the kind of behaviors that are welcomed. Curators might represent a small group of editors (a la a small hosted workshop). Or, by selecting *all* community members as curators, the community might behave as a democracy: one that intentionally includes active lurkers who visit often but vote less often, by predicting their opinions as well. Or, a community might establish a curation group that is a stakeholder roundtable [31], ensuring that minority voices are heard, unlike a straight majoritarian vote. By making explicit and transparent who is guiding decisions in a community, curation may facilitate deliberate discussion in each community about what sort of space they want to inhabit.

In summary, this paper contributes: (1) the Cura system to empower curation on large volumes of content without continuous manual review by drawing on community members' behavior; the system consists of both (2) a transformer-based curation model that predicts curators' opinions—whether they will upvote or downvote—based on community feedback, and (3) curation interface patterns such as frontstage and backstage spaces that allow community feedback to power the curation model's input; the model and interface are realized in (4) a functional implementation of these on the Reddit platform. Through this work, we aim to concretely demonstrate the feasibility of curation at social media scale.

[2]We use upvotes as the predicted curator behavior in this paper, but this action could be replaced with any other behavior: e.g., moderation decisions, likes, favorites, retweets, comments, or emoji reactions.

2 RELATED WORK

2.1 Curation, moderation, and norms in social computing

Curation, in a social computing context, denotes the process of a curator selecting a subset of the content that fits their preferences or the demand of a larger community, where the content can be news articles, user submitted posts, or comments [67, 78]. An early study of e-mail list moderators lists reports that they would sometimes engage in curation decisions about what content to allow, claiming that "quality must be maintained or the audience will desert the group" [5]. Such tensions persist today [81]. Many individual users proactively curate for themselves, e.g., by following certain users or groups [46], blocking certain sources [54], and searching [39]. Users productively curate for other users as well, e.g., by reframing and sharing content with designated groups of people [46, 50, 55]. Some users also curate to meet their self-expressive and self-representational needs [9, 77]. This self-curation is a critical part of an effective social media diet, though it has less of a direct influence on the wider community.

Curation for social media communities arises in some modern web contexts: Diakopoulos notes that a small subset of comments at the New York Times are flagged as 'NYT Picks' [21], and Bruns identifies that Kuro5hin and Plastic allow users to vote on submitted stories before they are published [7]. Even at a limited scale, these curation approaches require extensive human effort: in the case of the New York Times, it requires a dedicated paid team. Our objective is to develop a system that can support this sort of curation at scale without overwhelming or overly bottlenecking on human curators and the time that community members need to wait for approvals.

On the other hand, the term 'algorithmic curation' in previous literature mostly refers to content personalization for individual users based on machine learning algorithms [27, 75]. Such personalization algorithms create different experiences for each community member, so they are not community-level curation, and have also been implicated in increasing polarization and creating echo chambers [27]. In contrast to this prior work, our goal is to facilitate group-level curation, so our approach models a single decision criterion (the curator's taste) on behalf of the entire community.

The practice of upvoting and downvoting content by users and subsequently ranking it according to their votes on social media platforms such as Reddit can be viewed as a form of crowdsourced curation at a group level. However, it limits the pool of curators to the active members of the community. Therefore, our objective is to broaden the scope of curators and open up more possiblities for curation, for example to encompass lurkers, or to focus on a specific set of community members.

Curation is also related to the practice of moderation. Moderation is typically focused on the task of removing objectionable or norm-violating content: as Gillespie puts it, "both to protect one user from another, or one group from its antagonists, and to remove the offensive, vile, or illegal" [30, p. 5]. Much of the labor of moderators is focused on this sort of removal [47, 60]. Curation, as a form of moderation, can be separate from removal: as Wang and Diakopoulos put it, "highlighting high-quality content [is] a moderation strategy" [30, 78]. Inductive work has identified 'curator' as a metaphor that moderators may use to describe themselves [67]. Under the grammar of moderation [33], including principal techniques of moderation (nouns) and important distinctions in how moderation is carried out (adverbs), we can describe this existing curation practice as a manual, ex ante, centrally executed organization of content. In this paper, we draw on this practice to imagine a version of curation that can scale to larger communities, instead being described by the grammar as automatic, ex ante, and distributed in nature. Such curation has yet to be supported by a socio-technical system. Similar to those moderation tools that are designed for content deletion and removal at scale, we envision curation to be a collaborative effort of human curators with the support of algorithmic tools [30, 69].

Curation is not the only method of shaping norms in an online community. Community members respond to and replicate the behavior they see around them [13, 58], learning and following what are known as descriptive norms [41]. In broadcast, this makes violations visible to all members, and the violations tend to accrue engagement [37], creating a negative spiral in which more and more norm-violating behavior may arise over time [17]. Making norms salient can increase adherence amongst community members [18], for example by including permanent notes reminding members of the rules when they are authoring new content [49]. Identity signals can also influence norms, for example anonymity vs. pseudonymity [41] as well signaling ingroup identity on the platform [68]. By giving curators control over the ways in which designed signals are portrayed to the community, curation seeks to engage similar mechanisms.

2.2 Algorithms supporting social media

We draw on advances in machine learning and social computing to enable curation. Most machine learning classifiers only model one voice, the ground truth it has learned from the dataset, often reflecting the majority voice of the dataset labelers [30, 32, 61, 65]. Such a classifier can not satisfy the need for different communities with distinct tastes and values, and can result in the "tyranny of the majority" by ignoring minority groups' viewpoints [31]. Researchers have been calling on the development of algorithms that can balance the needs and values of multiple stakeholders and achieve collective goals (e.g., [1, 22, 51, 64, 71, 80, 83]). We draw inspiration from jury learning [31], which learns to predict how each individual user would label unseen examples to enable selection of which voice the community wants to listen to. We extend this concept to learn from other community members' votes in order to make a more accurate prediction.

AI models on social media such as Twitter and TikTok draw on insights from recommender systems to personalize content based on individual users' preferences. Our approach similarly make personalized predictions and draws on insights from recommender systems: we develop a user embedding as part of the curation model, allowing the model to make vote predictions for any user in its training data. Our model, however, is expected to update its prediction with each new vote from a member of the community. Researchers on recommender systems have articulated the need for value-sensitive design [16, 72], particularly to integrate human values and explicitly optimize the recommender systems for higher measurements on those values [72]. Our approach to this issue is to ask each community to make an explicit value statement through its decision of which members serve as curators, and which posts those curators upvote.

Collaborative filtering is a group of popular algorithms for building recommender systems which leverage the known preferences of a group of users to predict the unknown preferences of other users [59, 73]. Memory-based collaborative filtering, which calculates the similarity between items or users based on common users or items [43], is widely-deployed in commercial recommender systems [35, 48]. However, it suffers from sparsity issues when there are few common items, and is hard to scale for large datasets [73].

Particularly salient for our context is the cold start problem: collaborative filtering cannot make predictions on new items. Content-based recommender systems help overcome this problem by leveraging features of the users and items for making predictions, where the features include categorical features like genre, content, and URL domain [43, 57, 73]. In our work, we draw on content features from natural language processing [70]. Such techniques include Long Short-Term Memory [34], or transformers [76] such as BERT [20], a pretrained model designed to understand natural language text. These architectures can be used to model categorical attributes of users and items, as well as model users' historical behaviors—a sequence of items the user has interacted with [74, 84]. We exploit this capability of BERT by directly inputting the features of the user, the community, and the post as tokens toward prediction.

Under curation, only posts that are celebrated by the curators are published, preventing the spread of norm-violating posts as well as the resulting harm. In contrast to many algorithic moderation tools which directly measure the content toxicity, an algorithmic curation tool would attempt to forecast the reception of content by a particular curator or group of curators. Researchers have already developed models for forecasting conversational events—what a piece of content might bring up, whether it will be perceived as helpful [3] or lead to derailment [14, 66], steps toward the proactive, discerning application of curation that we envision.

3 CURA: DESIGN AND INTERACTION

In this section, we describe our design goals of the curation metaphor and the system, *Cura*, that we built to realize it.

3.1 Motivating Scenario

Several years ago, Alice created an online community about technology news because of her interests in security and regulation. It started as a group with people who shared her interests, but over time expanded to cover all tech news. Eventually, Alice and the original cohort of members left: the community no longer featured content that they cared about as the tides had turned towards crypto. Alice decides to restart a community around technology security and regulation news, but does not want to experience the same problematic shift again, so she opts to implement curation for this new community. She makes herself and some of the more active participants from the old community into the curators for this new one. This time around, when people join the community, they see news stories around recently discovered security vulnerabilities at the top of the feed. When someone tries to flood the community with thinly veiled cryptocurrency scams, the content remains backstage and not visible to most members, while a story around the SEC considering cryptocurrency regulation does move frontstage, since it fits Alice's vision for the group.

3.2 Design goals

Broadcast in social media prioritizes low friction to sharing content, roughly akin to giving everybody in the audience a microphone. Curation is more akin to a hosted show, where there is greater friction to getting content shared with the entire group, so the host can better shape the conversation. Each has its place in our social ecosystem: broadcast social media for more rapid-fire and free-flowing spaces, and curation for slightly slower spaces that pursue a particular set of norms.

Curation already occurs in some communities online, but it is limited by the sheer amount of effort required to pre-review every submission in large-scale communities. This project explores a design space in which curators loosen the reins slightly to allow an algorithm, with oversight, to aid them in curation. They do so by allowing the algorithm to learn from the curators' voting behavior and how it can be predicted by a combination of topic and votes from others in the community—which other community members' votes are predictive of the curator's votes? When the system gains confidence in the curators' likely behavior, it allows the post to move forward. To integrate this model into a system, we introduce interface approaches for curation, including a backstage for content that is not (yet) above the threshold for curation, as well as interfaces for articulating the curators for each community.

We detail here our design goals for our curation system:

(1) Operate on large-scale communities without overloading curators with work or grinding the community to a halt waiting for approvals. Moderators are overworked [60], and large communities cannot come close to reviewing each piece of content for norm violations. If members find that

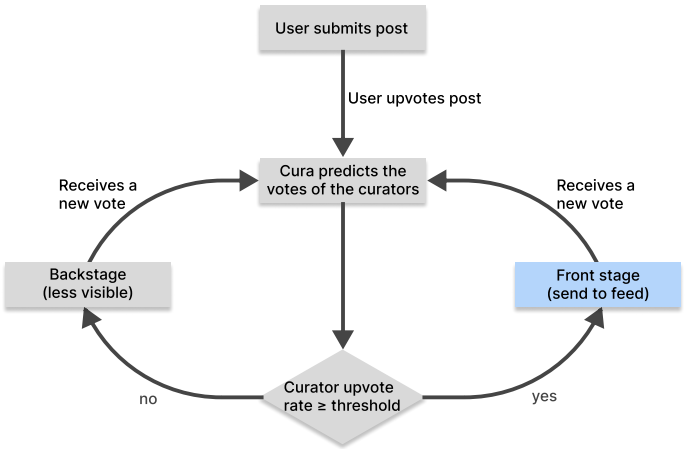

Fig. 2. A post's flow through Cura: When a new post arrives, Cura estimates the proportion of curators who would upvote it. If the proportion is above the community's curation threshold, the post goes live in the frontstage feed that is presented to members by default. If the proportion is below the community's curation threshold, it remains in a less visible backstage area which can only be viewed by trusted members or if members actively browse it. When a new vote comes in from a community member, Cura updates its estimates and reroutes the post to the frontstage or backstage appropriately.

their content is left indefinitely in a review queue, they may stop posting. So, we must thread the needle by identifying content while requiring manageable oversight from curators and only small delays. We achieve this by leveraging active community members and learning the correlations between the curators' opinions and those community members.

(2) Identify posts that curators support for their community. Toward empowering communities to better execute on their desired norms, curation models the curators' opinions on content in the community, and only promotes content that those curators would support. The design insight that motivates our system is that curator attention is scarce, while community attention is more plentiful. So, we draw on signals from community members while maintaining the curators' point of view as a North Star.

3.3 System interaction

We instantiate these ideas in *Cura*, a curation system for social media. Cura is currently implemented as an alternative feed visualization for Reddit, and is envisioned eventually as a standalone social media site. In the system, a community administrator selects a set of community members (Reddit accounts) as curators for their communities.[3] The interaction for this process is similar to selecting moderators. The administrator then selects a curation threshold that determines when a post gets selected to share with the group. For example, a curation threshold of 50% implies that 50% of the subreddit's curators must either manually vote to upvote the post, or be estimated to upvote the post, in order to share it. As the administrator selects curators, Cura supports them by visualizing information about the curator's history and activity levels (Figure 3). Members browsing the subreddit continue to upvote or downvote posts as before according to their own preferences, and

[3]Because our system is not directly integrated with Reddit, Cura does not directly enforce a requirement that it can only be used by administrators. We envision that, if this were a standalone social media platform, it would be available to community administrators.

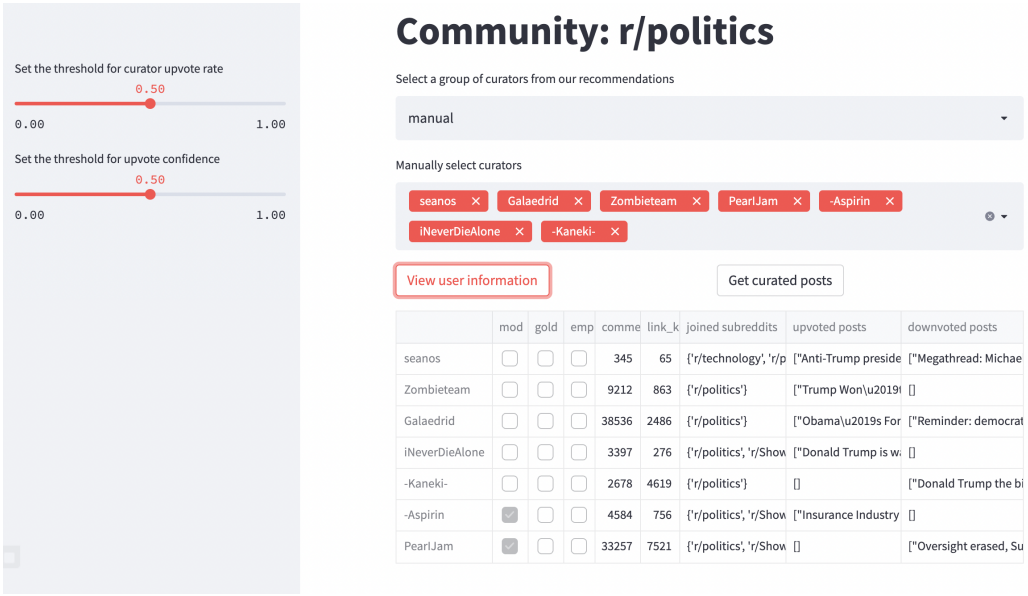

Fig. 3. Administrators define the curators on behalf of the community, set the threshold for how many curators should be estimated to upvote a post in order for it to go live to the entire community and the threshold for how much confidence should the estimation have to predict a curator upvote.

posters continue to submit posts as before. The key difference is that these upvotes and downvotes are now used to predict curator behavior: once the curators are defined for a community, Cura draws on signals from the post content, curator voting history, and upvotes and downvotes from community members to estimate whether each curator would upvote the content. If the curators' estimated approval clears a threshold, the post is shared in the community's public feed. Further votes can update the prediction, even hiding the post again if it falls below threshold. The system learns which community members are highly correlated with the curators and which are not, so it can ignore trolls, but update quickly with informative votes. Figure 2 illustrates the flow of the system. Below, we detail the different interfaces through which users interact with Cura.

Selecting curators. Curators are the members of each community whose upvote behavior is emulated for decision-making. In Figure 3, the administrator or moderator uses Cura to define curators for their community. Cura enables them to view information about community members, including the communities they have joined and posts they have upvoted or downvoted, and select a subset of users as curators. In addition, this administrator selects (1) a curation threshold determining the percent of curators who must support a post (through a combination of actual upvotes and the model's predictions) for the post to go live to the community, and (2) a confidence threshold above which Cura believes that a curator will support a post. The existence of two orthogonal thresholds exists primarily for expert use: for most cases, the first one suffices. The second threshold exists in case the administrator needs to make the model more or less conservative in its predictions—if it would be preferred, for example, for the model to only grant a curator's upvote when it is extremely certain. As an example of how combining these thresholds affords additional configurability for experts when needed, an administrator might state that the model must be 90% confident to predict that an individual curator will upvote in favor of curating a post into the community, and require at least half of curators to have voted in favor in order for the post to move ahead.

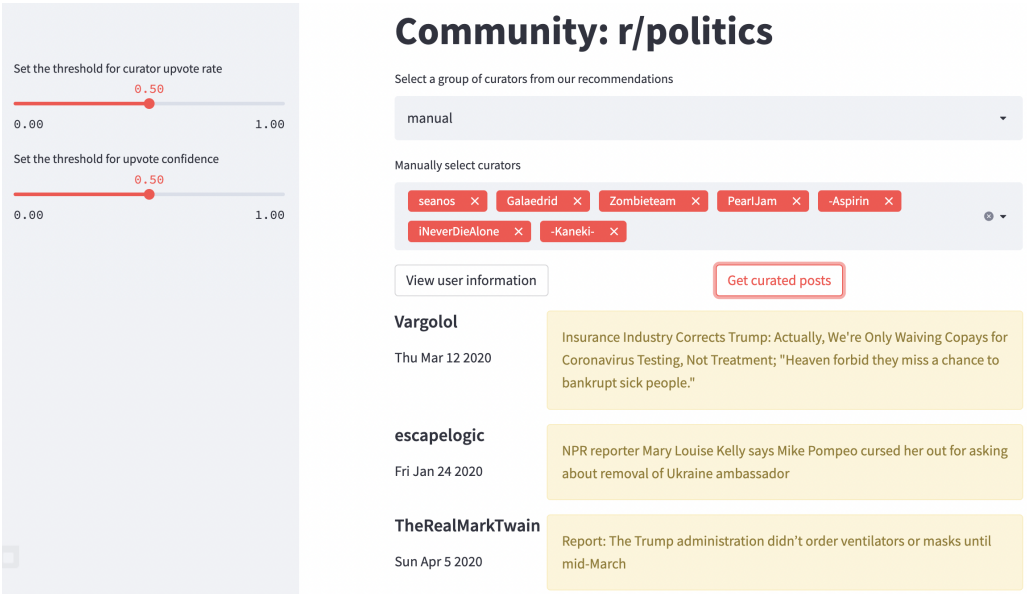

Fig. 4. Cura enables exploration of community feeds based on the selected curators and the set curation threshold. Here, Cura presents a version of Reddit's r/politics community that is filtered based on a set of seven curators.

How might a community select curators? Curators might be selected based on their interests and opinions, identity, or experience and past behaviors in the community. In Cura currently, we leave control of the curator selection to the community via the administrator, so that Cura can be guided by the community's values. We envision that curators might eventually be selected through election [82], named by trusted individuals such as administrators, or created de facto by the community purpose (e.g., consider the celebrity of a fan community). In the Discussion section, we reflect on how this selection process centralizes the power in the hands of the curators, and the governance processes that might aid it or prevent communities from coopting this approach for harm. In our design, curators are envisioned to be community members who are trusted by their peers, and selected based on factors such as expertise, opinion diversity, and identity diversity (Section 6.4).

By selecting curators, communities create different genres of community curation. A small group of selective curators might create a space with a highly specific purpose and strident sense of taste, and a more democratic community might select every member in the community as curators: not just the most active contributors, but also those who are often present but rarely vote (e.g., lurkers). The only technical requirement is that curators must be active enough for Cura to learn patterns in their behavior, e.g., at least 5 votes in the community. In the absence of this minimum engagement, there exists a potential risk of low-quality user modeling.

Generating the community's feed. We refer to the posts that clear the curation threshold as the community's *feed* (Figure 4). We recommend setting the curation threshold as 50%, so that a post needs to receive a majority of upvotes from the curators to be featured in the feed. The administrator can experiment with different thresholds and evaluate their impact on the community's feed through the interface (Figure 4). Subsequently, the curation threshold can be dynamically adjusted to align with the administrator's preferences.

r/Feminism members curate r/Teenagers	r/MensRights members curate r/Teenagers	r/LesbianActually members curate r/Teenagers	r/gay members curate r/Teenagers	Random members of r/Teenagers curate r/Teenagers
Some girl bragged about having a relationship for 5 months. Jokes on her! I've been with loneliness for nearly 17 years	My crush was complaining about not having a boyfriend and then I said "I could be your boyfriend" She laughed	Your cat isn't a "Thicc Boi" it's fucking overweight	No one fucking loves me my parents are abusive and i constantly get fucking bullied because i am "the quiet kid"	Ok so who's joining?
PSA: This subreddit is not a dating app, you horny teens	I think there's a squatter in my attic	iCarly was the first e girl. prove me wrong	Just so proud	Being the smart kid is a stress
Happy Alentines Ay for those of us not getting the V or D	For all you teens out there that are in "that" mood, here you go	So does that mean I'm in a relationship?	Those bastards lied to me	Why do they always scream.
My boyfriend broke up with me, I got a 72 on an exam, and my dad had a heart attack. Happy Valentine's Day!	My country is fucked.(UK) Government says 60% of the population needs to be infected with the corona virus to develop hers immunity	i am very fancy	R.I.P kobe. He was my legend, he will always live on in my heart. Still cant belive he actuelly died, so sad :(	He is now known as Steve from Accountin

Table 1. Feeds from the r/teenagers subreddit community, created with Cura by changing the set of curators for r/teenagers. Selected posts were those only preferred by one group of curators or that ranked substantially higher in one curated feed than others by curator upvote rate.

Table 1 demonstrates several different feeds curated from the r/teenagers subreddit community, where the curators of the first four feeds are the intersection of the active users in r/teenagers and the active users in r/Feminism, r/MensRights, r/LesbianActually and r/gay, respectively. The fifth group of curators is 50 randomly chosen users in r/teenagers. The feeds can be strikingly different from each other, with the r/teenagers feed curated by users in r/Feminism and r/LesbianActually promoting more woman-centric viewpoints. Choosing to empower a particular set of curators effectively also means choosing to diminish the influence of conflicting community members: the r/MensRights feed is clearly different from that curated by r/Feminism, for example, but both of these are viewpoints represented today within the broadcast r/teenagers community. Curation gives a community the chance to choose whose voices they want in control when there are conflicting values.

Cura's curation threshold determines the percent of curators who must be estimated to upvote the content or actually upvote the content. Table 2 demonstrates the r/Teenagers feed curated by all community members who have voted ≥ 5 times under increasingly strict thresholds. As the threshold increases, the content is more pro-social, but the risk increases of a false negative in the model prediction. Cura enables the practitioner to iteratively explore different compositions of curators and different thresholds to see how they will influence the resulting feed (Figure 3).

r/Teenagers curation threshold of 80%	r/Teenagers curation threshold of 60%	r/Teenagers curation threshold of 40%	r/Teenagers curation threshold of 20%
For the two people who wanted it! An update to my artwork	My cat hurt his foot a few days ago. To lift his spirits, let's make him into a meme format.	Almost lost my cool there	Any hot grills wanna have a chat
My dog Teddy got put down today, so I drew him as a last goodbye	Like I did the best mom,give me my ps4 back	Boys who don't like them selves and think their ugly. If you would like to answer, why?	can someone please crush my skull with a hammer? im having delusions about everything
You dare use my own spells against me, Potter?	saw this kid at the cafeteria browsing r/teenagers, make him see himself	Too many virgins on this sub reddit lol	Just chatting with the GFs sister
Same mirror, but one year apart! Anorexia recovery has been hard but seeing these pictures makes me proud of how far I've come	A doctor just flirted with me today. She told me that I am too sweet. Well her exact words were "severely diabetic" but I know what she meant.	Unpopular Opinion: heterophobic. Nobody is born straight. Strait people shouldn't adopt because straight people raise straight kids.	Driving tip for all you newbies

Table 2. Curated feed for the r/teenagers subreddit community with every community member being a curator. Each column shows post submissions with the curator upvote rate above a certain threshold but under a higher threshold, i.e., post submissions in the columns to the left are also above the threshold for this column.

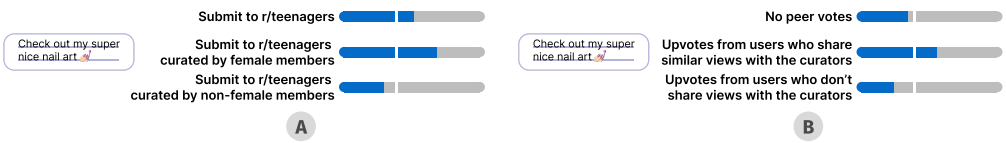

Fig. 5. Conceptual visualizations of how the same post might fare via curation across different communities and different voting groups. (A) Changing the set of curators affects whether posts on teen girls' and womens' issues clear the curation threshold. (B) If a post is upvoted by members who have not historically shared opinions with the curators, it is not curated as above threshold into the community.

An end-user visualization displays how close each post is to clearing the curation threshold for that space (Figure 5A). Given the same number of upvotes, the outcome may differ depending on which community members upvoted the post (Figure 5B).

Backstage. We introduce a backstage in Cura in support of content that is not yet above the curation threshold. If integrated into existing platforms such as Reddit that keep all posts visible, an implementation of Cura as described so far could use the curation estimate from 0% to 100% as a ranking score. This would enable a version of curation that integrates into a feed ranking algorithm, or be translated into a curator upvote visualization (Figure 5).

However, as a standalone platform, our instantiation of curation in Cura enables a split of content between the *backstage* and *frontstage*. Posts live in the backstage before the curation algorithm is confident in their quality, where community members may manually visit and vote to update the prediction; by default, however, members see the curated frontstage feed. When the algorithm

updates its prediction for the curators to be above the chosen threshold, posts move to the frontstage where the whole community can see them.

Backstaging enables community input—so the curators do not need to manually review every post—while keeping norm-violating content and anti-social behaviors in a low status and less visible area, mitigating potential damage to the community from norm degradation and harassment. If less visibility is desired, the community administrator could only give access to the backstage to trusted community members. Just as Reddit identifies promising content based on a few users actively browsing the "new" feed, a standalone version of curation social media such as Cura will rely on a few community members exploring the backstage and upvoting content to bring it close to or above the curation threshold. Submitting a post to a community operating with Cura is akin to submitting a paper to a conference, where some posts may never make it to the frontstage. However, users have the option of submitting posts that align more closely with curators' preferences, increasing the chances of appearing on the frontstage.

4 TECHNICAL APPROACH

Our goal is to enable curation at social media scale. To achieve this, we do not require that curators manually review every post. We thus seek to model curators' votes based on other members' votes—in other words, predicting "If members X and Y upvoted the content, and member Z downvoted it, what is the curator likely to say?" We also require an infrastructure that can update its prediction after each new peer vote, so that content can dynamically move between the frontstage and backstage.

To achieve these goals, we present a transformer-based model, which we call a *curation model*. Our curation model builds on the recent success of transformers [76] to create a transformer to classify whether the target user will upvote a post. The input to the model is a concatenation of the target user (curator) and stringified features of the post, including the post text content, post author, community where the post is published, voting time, whether the post is flagged as NSFW,[4] and the URL domain linked by the post. Following prior work [4, 23], we use special tokens before each feature to indicate the type of the feature being encoded. We represent each user as a unique token so that they have a unique embedding in the model. For example, to predict at "Nov 6, 2023", whether user a will upvote "Glaciers in Europe are experiencing the most severe melting on record" submitted by the user b in a community named "News", the converted model input string will be "`[USERNAME]` [a] `[AUTHOR]` [b] `[COMMUNITY] News [CREATED_TIME] Wed Nov 6, 2023 [NSFW] false [SUBMISSION_URL_DOMAIN] www.washingtonpost.com [SUBMISSION_TEXT] Glaciers in Europe are experiencing the most severe melting on record`".

Our implementation of this curation model uses a BERT encoder [20] with a linear projector for binary classification. Encoding the input with the BERT encoder, the linear projector takes the encoding of the target user token as input and outputs a scalar p, which represents the probability of the target user upvoting on the post. p indicates the confidence of the model prediction. We obtain the final decision—whether the target user will upvote—by thresholding the probability p, e.g., at a default value of 0.5. Users can adjust the threshold to control the composition of the front-stage content (Figure 3).

The curation model draws on others' voting behavior to estimate the curator's vote. To achieve this, we train on history votes from peers in the communities[5], including the initial upvote from

[4] "Not Safe For Work", a flag used to label content that contains explicit or offensive material that may be inappropriate for a workplace or public setting.

[5] Votes on content from all communities contribute to the model training and prediction. While different communities may have varying voting norms, the curation model is capable of learning how users' voting behavior differs from one community to another based on the community name included in the model input.

the author themselves (Section 3.3). The model receives training supervision through a series of training examples, for example, that Users X and Y upvoted post P, User Z downvoted post P, and the curator upvoted post P. From this, it learns to leverage prior upvotes or downvotes in its training data via the overlapping signals on post P to predict the vote for a given user. Then, to make a prediction informed by current upvotes, when the Cura system receives a new vote on a post, we finetune the full curation model (including the BERT encoder and the linear projector) one step further on the new vote, then re-predict for each curator. This finetuning adjusts the model's prediction of the curator's opinion, creating a system that is responsive to community feedback. Additionally, it allows the model to place greater emphasis on recent votes over older ones, which facilitates the adaptation of the model to changes in users' opinions. Even if the model initially makes incorrect predictions on users' changing opinions, it can learn from users' votes and capture their latest preferences over time. Finetuning is rapid enough for interactive purposes with sub-second speeds; if faster speeds are needed, parts of the model could be frozen during finetuning.

4.1 Dataset

For training and evaluation, we draw on a dataset of Reddit upvotes and downvotes from accounts that voluntarily opted in to make their vote data public. On Reddit, users could navigate to their profile and opt in to make their voting history public, including the posts they voted on and whether they upvoted or downvoted the post.[6] We use the following post features provided by the Reddit dataset to compose the model input: unique Reddit username, voting time, subreddit, post author Reddit username, and whether the post is flagged as NSFW. We additionally use PRAW,[8] the Python Reddit API wrapper, to retrieve the URL domain linked by the post (if present[9]), as well as the text content of the post.

Due to computational resource limits, in this paper we only use a subset of this full vote Reddit dataset. We sample every vote on posts in a set of popular subreddits on the topics of politics (r/politics, r/Conservative, r/Liberal, r/Republican, r/democrats, r/VoteBlue), jokes (r/Jokes, r/Showerthoughts), science (r/science, r/ScienceFacts, r/technology, r/shittyaskscience), and gender, sexual orientation, and identity (r/Feminism, r/MensRights, r/LesbianActually, r/gay, r/trans) and lifestyle (r/teenagers).[10]

This sample results in 1,966,122 votes, which comprise 4.4% of the dataset, containing 518,798 different posts. We randomly split the dataset (votes) into training and test sets at a 80%–20% ratio. This voting data, collected from the real-world Reddit behavior, is very unbalanced: 74% of the votes are upvotes and only 26% are downvotes, and 10% of the most active users account for 86% of the votes while most of the other users vote rarely and are more typically lurkers. Our model design and training must account for these imbalances.

[6]This preference is visible on the "old Reddit" profile page, and still functions for users who click through to it. The preference is called "make my votes public". The newer Reddit design de-prioritizes this preference. The resulting dataset is at: https://www.reddit.com/r/help/comments/8x11lp/how_to_make_upvoted_public/ We utilize a dataset that collected these opt-in users' history voting data.[7]

[8]http://praw.readthedocs.io

[9]Only posts of the "link" type contain a retrievable URL domain. Other post types will result in an empty string.

[10]These subreddits were chosen due to their representation of distinct identities, interests, and views. The selection allows us to conduct experiments, as outlined in Section 5.2, whereby we designate users from certain subreddits as curators. We aim to confirm that the curation outcomes differ significantly depending on the curators selected.

4.2 Training

We use an implementation of the BERT transformer model from Huggingface [79]. We initialize the weights of the BERT encoder with the pre-trained BERT-mini [6], then expand the model vocabulary and token embedding by adding special indicator tokens for each feature and adding a unique token for each user. The embedding of the added tokens and the weights of the linear projector are initialized randomly. We then train our model to predict user votes by fine-tuning on our Reddit dataset.

To address the problem of unbalanced voting data, we apply a weighted binary cross entropy loss. For a user-post-vote triplet where user $a \in U$ voted $x \in V = \{\text{upvote, downvote}\}$ on post $s \in P$, where U is the set of users and P is the set of posts, we assign the weight to be

$$W(a, s, x) = \frac{1}{|\{(u, p, v)|u = a, v = x\}|} * \frac{|\{(u, p, v)|p = s\}|}{|\{(u, p, v)|p = s, v = x\}|} * w(x)$$

where upvotes receive $w(x)$ of 1, and downvotes 1.5. The weight is inversely proportional to the number of up- and downvotes the user has made, and inversely proportional to the proportion of users that have the same opinion x on post s as the target user a. This weighting ensures that the model prediction is not biased towards the most active users or toward the majority opinion. We add an additional weight for downvotes to balance the model's strong tendency of predicting upvotes.

We trained our model on one Tesla V100 GPU for 10 epochs with a batch size of 64 and a maximum text length of 512 tokens. We use Adam [42] optimizer with a learning rate of $3e-5$. When the Cura system receives new votes, we adopt a learning rate of $3.6e-5$ for finetuning. The hyperparameters are selected after a small grid search.

5 EVALUATION

This paper proposes that the Cura system can reshape the contents of a community, and that curators' opinions can be estimated effectively based on community members' opinions. In this evaluation, we test these proposals by mapping them onto two main questions: (1) Does our model estimate curators' votes accurately? (2) What impact does curator selection have on the algorithm's selected posts for a community?

5.1 Estimating Curators' Votes

Whether Cura is able to accurately predict curators' opinions on posts is a core question and is the cornerstone of other claims. To answer this question, we test whether our model can accurately predict votes from arbitrary users—essentially, any curator we might select—in the test set of our Reddit vote dataset. We use upvotes as proxies for content that these users actively want to see in the community, so our goal is to predict upvotes accurately.

5.1.1 Performance across users. To calibrate performance measures, we sampled a subset of upvotes from the original test set to form a balanced test set, where the number of upvotes is equal to the number of downvotes for each subreddit, so that random guessing is 50%. The balanced test set consists of 127,528 posts, with 103,818 upvotes and 103,818 downvotes in total. Each post corresponds to 1.63 votes on average, ranging from one vote to 52 votes, with a standard deviation of 1.88 and a median of one vote. Perfect accuracy of 100% is not feasible, given that people often hold self-inconsistent opinions and would not react the same way if they saw the same post a second time [10, 32].

Our model achieves an accuracy of 81.96%, with a ROC AUC score of 0.8903. For comparison, we trained an ablated baseline model that predicts the majority vote for each post. When testing

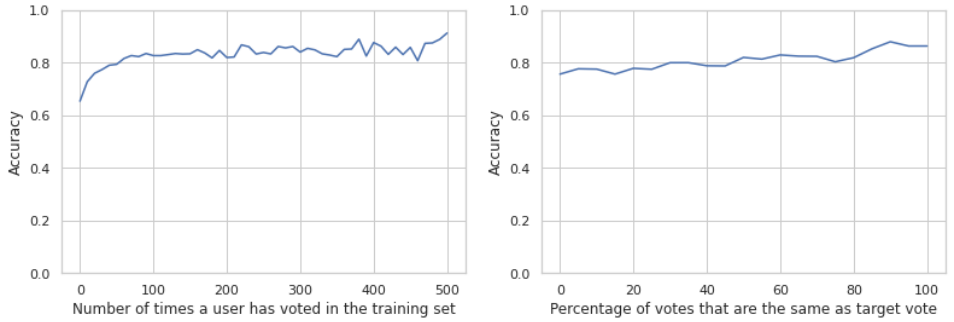

Fig. 6. (A) For active users who vote more than fifty times, the model achieves 80% accuracy. The model requires few votes to begin modeling users accurately, improving from 62% for users who have voted only once to 73% for users who have voted ten times. (B) Accuracy for target votes that are the same with different proportions of votes: the accuracy is always above 70% no matter whether the vote is the majority or minority opinion.

Actual/Predicted	Upvote	Downvote
Upvote	85039 (0.8191)	18779 (0.1809)
Downvote	18673 (0.1799)	85145 (0.8201)

Table 3. Confusion matrix of the vote prediction result on the balanced test set. The model achieves a high accuracy for both upvotes and downvotes.

on the same balanced test set, this baseline achieves 65.96% accuracy and a ROC AUC of 0.7241, indicating that our model better predicts curator votes than a majority vote.

Curators are likely active users, and the model performs at over 80% accuracy for active users. However, some communities might want to implement democracy as a curation strategy, which would also require estimating lurkers' opinions so that it can proxy their votes as well. How well does the model estimate lurkers' votes? Accuracy begins at 62% for users with only one vote in the training data, and raises to 73% after ten votes (Figure 6(A)).

Does the model actually learn the preferences of individual users, or is it copying majority vote? The model estimates both majority ($\geq$ 50% same votes) and minority ($\leq$ 50% same votes) opinions with roughly equal accuracy (Figure 6(B)). This indicates that the model is learning individual opinions rather than "piling on" majority votes.

5.1.2 Performance across topics. Does the model perform worse in certain communities? Table 4 reports accuracy across each of the subreddits in our balanced test set. Performance generally is stable, with a mean prediction accuracy of 77.86% (σ = 8.43%). The only low-performance outlier, r/ScienceFacts, has only 137 datapoints in the full dataset (108 datapoints in the training set)—too few to model effectively. Based on these results, we suggest having at least 2000 datapoints in a community's training set to ensure effective model performance.

5.1.3 Influence of peer votes. Peer votes should be able to influence the prediction confidence and even correct previously incorrect predictions. Even if the model estimates that a curator is not likely to upvote a piece of content, if several members whose opinions are correlated with the curator's opinion begin upvoting the content, the curator's estimated behavior should shift. To test this effect, we trained a model using the same method from Section 4.2 except that the training set

	r/politics	r/Conservative	r/Republican	r/Liberal	r/democrats
Number of datapoints	1,082,006	63,132	7245	2076	6890
Accuracy	84.49%	91.51%	88.34%	79.10%	85.10%
	r/VoteBlue	r/Showerthoughts	r/Jokes	r/science	r/ScienceFacts
Number of datapoints	7732	252,455	138,451	111,857	137
Accuracy	78.04%	78.46%	77.01%	76.31%	50.00%
	r/technology	r/shittyaskscience	r/Feminism	r/MensRights	r/gay
Number of datapoints	102,203	7992	5482	24,682	4257
Accuracy	77.39%	76.67%	74.20%	79.12%	77.52%
	r/trans	r/LesbianActually	r/teenagers		
Number of datapoints	2597	2483	144,444		
Accuracy	69.35%	81.76%	77.58%		

Table 4. The model performs similarly across a wide variety of subreddits with different number of datapoints in the full dataset.

and test set are divided so that there are no overlapping posts between the training set and the test set. Then, we randomly select one user-post-vote triplet in the test set as the target, randomly shuffle all other votes on the post in the test set and then feed those peer votes to the model one by one. For each new peer vote, we finetune the pre-trained curation model one step further, then evaluate the model prediction on the target vote.

As the model receives more and more votes from other users, both prediction confidence and accuracy increase (Figure 7): after receiving 10 peer votes, the average confidence of accurate predictions increases from 0.802 to 0.854, and the prediction accuracy increases from 73.03% to 82.19%. The accuracy further goes up to 92.02% after 20 peer votes. This increment is not a result of copying the majority vote from peers (Figure 7C): the model insists on its own prediction when it considers the target user to hold a different view from the majority.

A second test: the model should update its prediction when users with similar preferences as the curator vote. We constructed two fake peer vote datasets: we randomly select one user-post-vote triplet in the test set as the target, and then for that user, find the top 50 other users whose voting vector (vector composed by votes on the post in the training set) is most similar to the target user via cosine similarity. We use these voters to construct two datasets where those similar users vote (1) the same or (2) the opposite as the target vote.

Accuracy and the confidence in predictions increase more rapidly as a function of the number of votes when the votes come from similar peers (Figure 8(A) vs. Figure 7). The accuracy increasing quickly indicates that the model adapts to peer votes quickly when the peers hold similar tastes. Conversely, if the peers are voting differently, because these peers are especially similar to the curator, the model loses confidence and eventually flips its prediction to the opposite of the ground truth, dropping prediction accuracy to nearly zero (Figure 8(B))—as desired. However, since the peer votes contradict the model's known preferences of the users, the model is more reluctant to change its prediction than in Figure 8(A).

5.2 Different Curators Result in Different Feeds

Given the same inventory of post submissions, do different compositions of curators result in different algorithmically selected feeds? To answer this question, we compare curation results between different groups of curators.

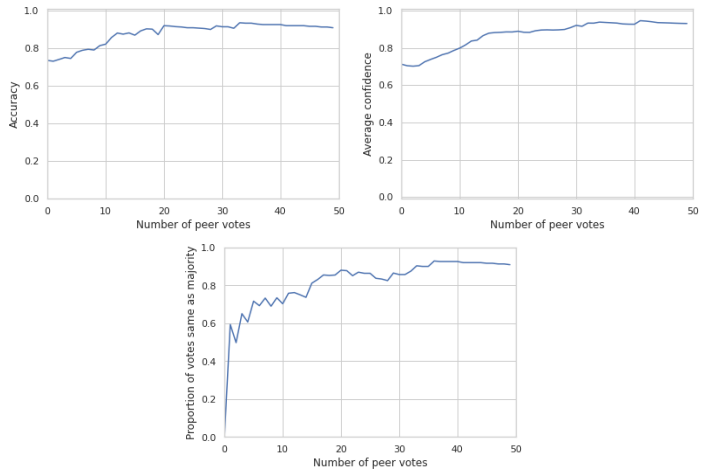

Fig. 7. (A) Accuracy increases rapidly as more peer votes are provided, and plateaus around twelve votes. (B) Confidence also increases as more peer votes are provided. (C) The model's estimated vote for a curator reacts as more and more peers vote.

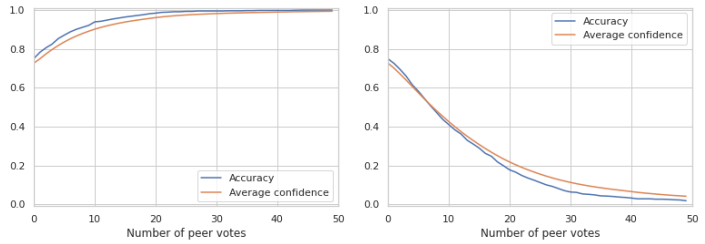

Fig. 8. (A) When peers with known similar preferences vote, the model adjusts its estimate of the curator's opinion quickly. (B) When peers with known similar preferences vote against the expected outcome, the model loses confidence in its initial vote and eventually flips its outcome.

Case study: r/politics. We mixed the posts from the subreddits "r/politics", "r/Conservative", "r/Liberal", "r/Republican", "r/democrats" and "r/VoteBlue" by randomly sampling 500 posts from each subreddit to create a super-politics community with a large inventory of politics-related posts that reflect different points of view and norms. The users of this community are the union of the users of the individual subreddits.

To select curators representing different points of view, we selected users who consistently upvoted posts in each of the constituent subreddits (except the more general r/politics community), specifically those who have voted ≥ 5 times in that subreddit and have a $\geq 70\%$ upvote rate for that subreddit. We also randomly sampled users from the supercommunity as a control group, thus obtaining 6 groups of curators in total.[11] We then applied those six sets of curators to the content of the entire super-community with the curation threshold for curator upvote rate being 50%, producing six different feeds.

In terms of content decisions, the preferences of right-leaning curator groups largely differ from those of the left-leaning groups (Table 5), while random users are in between but closer to the

[11]Based on this method, groups may have overlapping curators.

	r/Conservative	r/Liberal	r/Republican	r/democrats	r/VoteBlue	Random users
r/Conservative	1.					
r/Liberal	-0.342	1.				
r/Republican	0.837	-0.013	1.			
r/democrats	-0.259	0.957	0.052	1.		
r/VoteBlue	-0.301	0.952	-0.010	0.975	1.	
Random users	0.004	0.878	0.308	0.889	0.865	1

Table 5. Curators with different perspectives result in feeds with very different content. This table reports Pearson correlation coefficients of the curator upvote rates on individual posts, across groups of curators from different subreddits.

left-leaning groups, given that r/politics is left-leaning in general. We also present samples of the distinctive posts preferred by each group in Table 6: they strongly align with the taste of the curators.

Case study: r/teenagers. We next replicate the same result using communities that are more identity based rather than political. We randomly sample 500 posts from the popular subreddit r/teenagers to create the initial inventory of posts for a teenager community. Following the same method as before, we selected different sets of curators as users who upvote content both in r/teenagers as well as in r/Feminism, r/MensRights, r/LesbianActually or r/gay. We also randomly select 50 users from r/teenagers as curators, thus creating four groups of curators. Pearson correlations of curator post upvotes across these groups ranged from 0.614 (e.g., r/gay and r/LesbianActually) to 0.837 (r/gay and r/Feminism). Table 1 reports posts uniquely preferred by each group.

Case study: r/technology. In a final case study, we aimed to create a technology community that has a strong preference for posts about specific topics or posts with specific styles and formats. We initialized the technology community with 500 random posts sampled from the subreddit r/technology. We then selected four groups of users based on unsupervised clustering of vote patterns, which we label as: (1) Users who prefer human-centered content, e.g. technology that benefits people's health or can be applied to daily life, or labor rights in large technology companies; (2) Users who prefer content related to security and regulation, e.g. cybersecurity and software security, government guidance and misinformation; (3) Users who have broad interests, i.e., users who upvoted on a wide variety of topics; (4) Users who prefer user-generated content (e.g. discussion posts), instead of posting a link to other technology media. In Table 7, we observe that content curated by each curator group again strongly aligns with the tastes of the curators.

5.3 Online experiment

Does curation produce shifts that are strong enough to be recognizable and distinguishable from current community approaches?

To answer this question, we recruited $N = 20$ participants from Amazon Mechanical Turk with a Masters qualification who had a 97% approval rate and at least 5000 approved HITs, and paid them $7.50 each for a thirty-minute study. Participants were presented with fifteen pairs of feeds. Each pair presented two versions of a feed for a Reddit subreddit, one of which was curated by a target group, and participants' goal was to identify which of the two versions was curated by

r/politics curated by r/Conservative users	r/politics curated by r/Liberal users	r/politics curated by r/Republican users	r/politics curated by r/democrats users	r/politics curated by r/VoteBlue users	r/politics curated by random users
Nashville's Convention Center Discriminates	The Navy Has Decided To Restore Capt. Brett Crozier	Go Back!	Obama says White House response to coronavirus has been 'absolute chaotic disaster'	Former White House ethics director accuses Trump of 'laying groundwork' to interfere with presidential election	Emotional Schiff Speech Goes Viral, Delighting the Left and Enraging the Right
"If China knew of the outbreak way before they told WHO and covered up things, this is a crime against humanity" - says Italian politician Matteo Salvini	Fox Business Blaming Stock Market Drop on Sanders Is a Sign of Things to Come—Ignore the possible pandemic, blame the Bern	WHO officials praise US leaders on coronavirus pandemic response: Trump is doing 'all he can'	Trump: Presidents Have Infinite Power but Cannot and Should Not Ever Be Relied On to Do Anything	Health Care Workers Stand Up To Anti-Lockdown Protesters In North Carolina	'They Are Saving Our Lives': Demand Grows for Grocery Store Employees, Other Frontline Workers to Receive Hazard Pay Amid Coronavirus Outbreak
Nikki Haley backs investigation of WHO's COVID-19 response: America deserves answers	'They Are Saving Our Lives': Demand Grows for Grocery Store Employees, Other Frontline Workers to Receive Hazard Pay Amid Coronavirus Outbreak	Cuomo: You know who's doing a good job on NY's COVID-19 outbreak? Trump	The Navy Has Decided To Restore Capt. Brett Crozier	America doesn't want another Tea Party - Don't let Fox News fool you. 81% of Americans do not share the views of anti-quarantine protesters.	Sen. Tammy Duckworth spends Veterans Day in Mexico to be with veterans deported by the Trump administration

Table 6. Example posts from the politics community preferred by different groups of curators: either only curated by one group, or ranking much higher in one group than in others.

the target group. The correct feed was curated[12] by the stated group, and the incorrect feed was generated either by typical community upvoting (broadcast), or through curation by a different group. Example comparisons included: r/technology vs. r/technology curated by members who are also members of r/programming, r/Jokes vs. r/Jokes curated by members who are also members of r/teenagers, and r/worldnews curated by members who are also members of r/india vs. r/worldnews

[12]To enable more diversified curator selection for the experiment, we trained our curation model on a larger dataset. The larger dataset includes every vote on posts in the following set of popular subreddits on the topics of politics (r/politics, r/PoliticalDiscussion, r/Conservative, r/Liberal, r/Republican, r/democrats, r/VoteBlue), jokes (r/Jokes, r/Showerthoughts), science (r/science, r/ScienceFacts, r/technology, r/shittyaskscience), interest and ideologies (r/gaming, r/tattoos, r/MakeupAddiction, r/Music, r/punk, r/Fitness, r/travel, r/programming, r/hacking, r/Bitcoin, r/conspiracy, r/Futurology), movies and genres (r/movies, r/RomanceBooks, r/Marvel, r/marvelstudios, r/scifi, r/sciencefiction, r/Drama, r/anime, r/Documentaries, r/StarWars), lifestyle (r/teenagers), world, countries and regions (r/worldnews, r/britishproblems, r/europe, r/france, r/unitedkingdom, r/canada, r/australia, r/india, r/Philippines), religions (r/atheism, r/Christianity, r/Buddhism, r/islam), and gender, sexual orientation, and identity (r/Feminism, r/MensRights, r/LesbianActually, r/gay, r/trans).

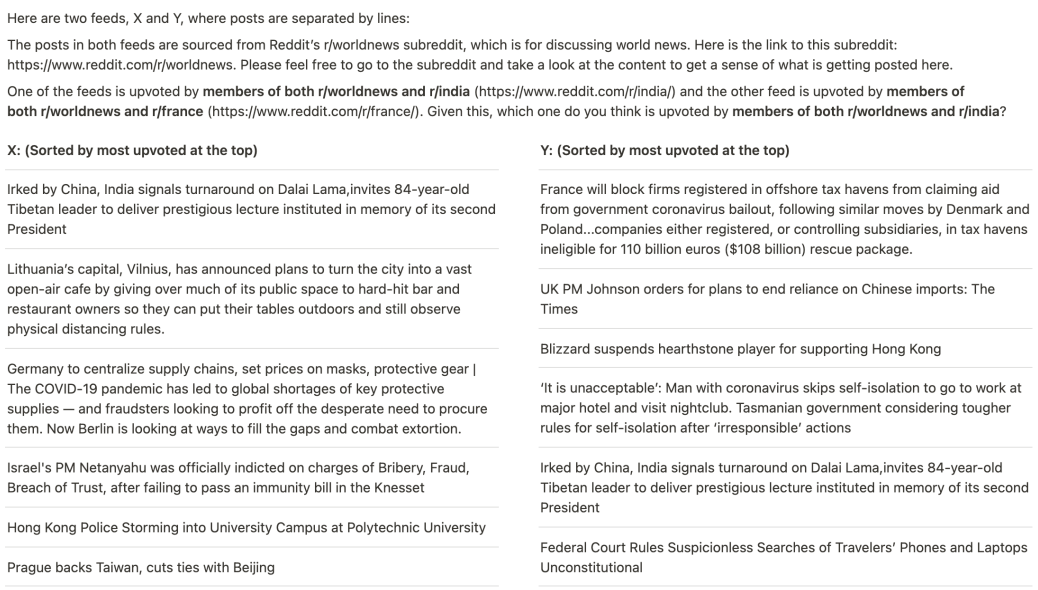

Fig. 9. An example pair of feeds for comparison in the online experiment. To avoid confusion that might stem from a description of the curation algorithm, we asked participants to select the feed "upvoted" by a target group of community members.

curated by members who are also members of r/france. Both pairs of feeds were generated by sampling the same five hundred posts from the subreddit, then ranking. Any traditional upvote (broadcast) feeds were ranked by the posts' actual Reddit score (upvotes - downvotes), and any curated feeds were ranked by the number of upvotes predicted by the curation model. The threshold for both prediction confidence and curator upvote rate were set at 0.5. Feeds were limited to the top fifteen ranked items. Participants were blind to which version of the feed was which. To avoid confusing participants with a description of the curation algorithm, we asked participants to select which of the feeds was created via upvotes from the target group. For more information on the fifteen pairs of feeds, please refer to the Appendix A.

For each pair, participants were told the subreddit and the curators, and were asked to guess which of the pair was generated by the curators as well as provide a written justification for their choice. To ensure the participants fully understood our task, we presented the participants with a training example, as well as the correct choice, at the start of the task.

We observed that 84.7% of participants' selections were correct, far above a 50% random guessing baseline. A one-proportion z-test ($N = 20 \times 15 = 300$) confirms that participants recognized the curated feed at above random chance rate (z=9.486, $p < 0.001$). These results suggest that curated feeds successfully execute recognizable shifts to the content of the community.

Participants were most successful when the curation group had distinct enough taste, and enough submissions in the subreddit representing that taste, to be recognizable. For example, every participant distinguished r/worldnews curated by r/india from r/worldnews curated by r/france, mentioning that the resulting feed explicitly mentioned "India" multiple times. In contrast, one of the more challenging comparisons was r/worldnews vs. r/worldnews curated by r/Liberal, likely because Reddit skews liberal and so the r/worldnews feed already embeds this point of view as a result. Future work in Cura could thus visualize for community administrators and moderators

Human-centered	Security and regulation	Broad interests	User-generated content
Yes, Americans can opt-out of airport facial recognition — here's how	Google tracked his bike ride past a burglarized home. That made him a suspect.	With more than 50 million US subscribers, Netflix has finally surpassed cable TV	We need to make it clear to the FCC that we want uncapped Internet access, for innovation in an increasingly data dependent world and user protection.
A Worker in Amazon's New York Warehouse Has Died of the Coronavirus	Facebook Will Ban Protests That Defy Government 'Guidance' on Distancing	Trump calls for 6G cellular technology, because why the heck not	Hey guys, Eric from Netflix, letting you know we're joining reddit and others for 'Internet Slowdown' Day Sept. 10th to protect Net Neutrality.
Toshiba says its device tests for 13 cancer types with 99% accuracy from a single drop of blood	In 2020, Some Americans Will Vote On Their Phones. Is That The Future? - For decades, the cybersecurity community has had a consistent message: Mixing the Internet and voting is a horrendous idea.	Robocallers blasted Americans with 26.3 billion spam calls last year - Robocalls are up 46 percent from 2017	If Google, Facebook, Twitter, and Yahoo really want to raise awareness on SOPA, they should follow Wikipedia's idea and shut down their sites and services for the day. TL;DR Goo, FB, Twit, and Yaho should do a wiki.
A Device That 'Prints' New Skin Right Onto Burns Just Passed Another Animal Trial	Macs now twice as likely to get infected by adware than PCs, according to research	Alicia Keys using sealable pouches to lock up concert goers phones to have a 'phone free show'	Eight members of Congress that voted to kill broadband privacy are now leading the charge to kill Net Neutrality as well

Table 7. Distinctive posts from the technology community preferred by different groups of curators (either only preferred by one group of curators or ranks much higher in one curated feed than others by curator upvote rate).

how much the curators' taste is similar to or different than the submission pool, general upvote pool, and other metrics, to understand what role curation is or is not playing.

5.4 Curation reduces anti-social content

Moderation requires substantial effort from volunteers to maintain pro-social norms, yet still one in twenty posts on Reddit post moderation contains violations such as misogyny, personal attacks, or bigotry [56]. Here, we demonstrate that curation dramatically reduces the rate of anti-social content without requiring any additional effort from moderators. By selecting curators who do not upvote norm-violating content, such content is not curated into the community's feed.

We focus on the challenging context of r/teenagers, following the previous method of curator selection, creating conditions for broadcast (raw Reddit score), democratic curation (modeling all users who have at least five upvotes in r/teenagers), as well as four conditions curating r/teenagers via its users who also actively upvote content in r/trans, r/gay, r/LesbianActually, or r/Feminism. We sample 1500 random post submissions from r/teenagers and then measure the macro-norm violation rate of the subset of these posts curated in each condition with a curator vote threshold of 50% by replicating the AI+crowd worker annotation pipeline for identifying Reddit macro-norm violations[13] developed by Park et al. [56]. For the crowd annotation, we enlist United States-based Mechanical Turk workers with a Masters qualification who have 97% approval rate and at least 5000

[13]A post is regarded as norm-violating if it violates any of the eight macro-norms identified by [56], including using misogynistic or vulgar slurs, inflammatory political claims, bigotry, verbal attacks on Reddit or specific subreddits, posting pornographic links, personal attacks, abusing and criticizing moderators, or claiming the other person is too sensitive.

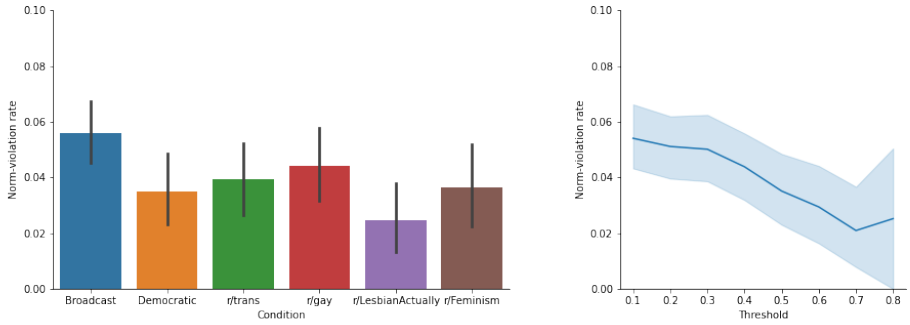

Fig. 10. (A) Norm-violation rate of broadcast (5.60% [4.49, 6.89]), democratic curation (3.51% [2.36, 5.0]) and multiple small group curation. (B) Under democratic curation, as the threshold for curator upvote rate increases from 0.1 to 0.8, the norm-violation rate decreases from 5.41% [4.31.6.68] to 2.52% [0.00, 5.43] (the small increase is due to the bias caused by few data points). The error bar represents the confidence interval.

approved HITs. Three workers review whether a particular post violated each of the macro-norms from prior work, and we label each post as a violation if two of the three workers labeled it as violating at least one macro-norm.

We first replicate prior work by finding that the broadcast condition features one in twenty posts violating at least one macronorm: 5.60%, 95% CI [4.49, 6.89]. The curated communities substantially reduce this rate, with no extra manual moderation effort. Democratic curation—the simplest curator selection—decreases the norm-violation rate to 3.51% [2.36, 5.0] (Figure 10(A)). Better results can be further achieved by selecting appropriate community members as the curators: the community curated by users who actively upvote in r/LesbianActually achieves a norm-violation rate of 2.48% [1.39, 4.05], reducing the overall rate by over half. A Chi-Square test confirms that the difference between conditions is significant ($\chi^2(5) = 13.4$, $p < .05$).

Setting different thresholds for the curator upvote rate results in different front-stage feeds as well (Table 2). High thresholds (e.g., 0.8 = 80% of curators upvoting a post in order to be curated) encourage positive posts and sharing achievements, 0.6 includes more humor and joke posts, and 0.2 shows weak control over content, allowing diverse values and negative posts. We again measure the norm-violation rate under democratic curation, this time varying the curation threshold, and find that the norm-violation rate significantly decreases as the threshold increases, down to again roughly 2% (Figure 10(B)), indicating that higher thresholds encourage tighter norm adherence and reduce harm.

5.5 From broadcast to lurker-inclusive democracies

Compared to broadcast, which only amplifies the voices of users who vote, communities can build more democratic spaces by selecting a wider set of users as the curators, or create a more centralized community by selecting a small group with distinctive taste. To evaluate the effect of switching to a more democratic community, we conduct an experiment on r/teenagers. The *broadcast* condition ranks by Reddit's raw score (upvotes-downvotes), which effectively only considers actively voting users. The *small group curation* condition follows the same method as before by selecting different sets of curators as users who upvote content both in r/teenagers and r/Feminism, r/LesbianActually or r/gay. Finally, the *democratic curation* condition includes all participants in the community with

	Broadcast	r/Feminism	r/LesbianActually	r/gay	Democratic
Broadcast	1.				
r/Feminism	0.503	1.			
r/LesbianActually	0.585	0.686	1.		
r/gay	0.611	0.760	0.772	1.	
Democratic	0.664	0.812	0.841	0.919	1.

Table 8. Different conditions result in content with very different ranking. This table reports Spearman's rank correlation coefficients of the feed ranking across broadcast, small group curation and democratic curation.

$\geq$ 5 votes in the community (i.e. those who are at minimally active and likely understand the community's norms) as curators.

We then measure the Spearman's rank correlation coefficient between the resulting ranked feeds (Table 8). Broadcast and democratic curation feature a rank correlation of 0.664, suggesting that post rankings might change substantially between Reddit today and a democratically curated system. Democratic curation combined the preferences of multiple small groups of curators: the rank correlation between democratic curation and curation by curators from r/Feminism, r/LesbianActually or r/gay are all above 0.81.

6 DISCUSSION

In this section, we reflect on our contributions and limitations. We discuss how communities might use the Cura system in practice, and discuss potential risks and ethical considerations of our approach.

6.1 Curation, broadcast, and the rules of social media platforms

Socio-technical systems shape the behavior of the people in them, and the systems are shaped by the people in them [2]. These systems embed rules that either directly control or indirectly influence who can speak, in what ways, and when.

Online community software today has largely aligned around the design metaphor of *broadcast*. In a broadcast metaphor, platforms focus on empowering users to share content instantaneously: sending posts or replies to all others in a community, to a newsfeed, or to all followers, before inspection. Research has examined how to encourage participation [8], shape newcomer behavior [49], and manage anti-social behavior [41] in systems with such a broadcast metaphor.

Curation sits in a space of alternative design metaphors for social media. Could social media provide more social translucence [28], stronger notions of consent [38], or other rules that are not tethered to or trying to fit within the broadcast metaphor [36]? We explore curation here because it inverts the logic of instant broadcast, providing trusted parties with more levers to shape the social space, and trades off content quantity for control. Future research can continue to explore this and other metaphors to offer a more varied palette of social designs for communities to select from.

6.2 Application in practice and implications for design

Cura is currently an algorithmic and visualization layer on top of Reddit data. Cura would require a community to change its back-end voting, or install a bot to do this on its behalf, to fully instantiate curation on existing platforms, not a small change. Still, in terms of the user experience of reading and upvoting posts, Cura is not far from existing social media platforms such as Reddit. In principle, Reddit or Facebook Groups could enable groups to nominate curators and then offer a separate ranking view for curation. Moreover, curation can also be combined with a personalized feed:

curation can serve as a common filter prior to personalized ranking. While curation determines what is the inventory of the posts publicly visible to the community, personalized feed ranking can determine the subset of the inventory presented to the end user.

Automated moderation algorithms offer another possible horizon for this approach. Current moderation algorithms focus on simple term lists or regular expressions, yet often falsely label appropriate content as inappropriate. It is possible that variants of this approach might offer an alternative that can generalize based on learnable model parameters, or for example the ability to predict whether curators will block a piece of content.

Social media platforms that are composed of multiple communities with different topics and tastes, e.g., Reddit, are the best fit for curation as described. For platforms without community structure, e.g., Twitter and Facebook, the implementation would need to be adjusted, perhaps by considering the whole platform as a global community, where curation can help to maintain the global norms in the platform. Alternatively, in the case of Twitter communities and Facebook Groups, where the platform has implemented communities for users to share information, ideas, and experiences related to a specific topic, Cura can be immediately applied. By selecting experienced members of the groups as curators, Cura can provide a more focused vision by selectively displaying curated discussion threads.

In practice, it may be more straightforward to launch a new community platform for curation. Doing so would help portray a simpler mental model to users, where it is clear that all communities are curated, and the voting user interface can be better customized to communicate curation thresholds.

More broadly, our work is a reaction to social computing policy and algorithmic solutions that assert that communities and platforms can only perform *downstream* reactions and clean-ups after anti-social behavior has occurred. As the social computing research community has demonstrated, however, a more effective approach may be to design *upstream* changes that shift behavior so that the anti-social behavior never arises in the first place. Our work serves as a viable alternative to these reactive solutions, and other design alternatives are also available. For instance, a platform may consist of hierarchically connected communities that enables high-quality and cross-cutting content to traverse the network.

There may be concerns regarding the potential negative impact of Cura on user engagement, which might further lead to degradation in model performance. However, we posit that users are still motivated to vote as before, given that their votes can improve the performance of the curation model and have an impact on the outcome. While there may be instances where a user's vote holds minimal weight due to behaviors such as trolling, the potential secondary effect that they vote less still bring about positive benefits for the community.

6.3 Scaling from small to large communities

Cura is not a good fit for every community. As a general gideline, for a small, tightknit community that contains less than 50 people who know and highly trust each other, little curation may be necessary—the community may be a better fit for broadcast. In this situation, the curation model may also be less accurate, given less voting data for model training. If the community grows, however, this can provide valuable training data for the curation model and allow the community to engage with curation immediately. For medium-sized communities consisting of 50 to 500 users, we recommend utilizing Cura. In the case of large communities with over 500 users that desire a curated flavor, we strongly recommend adopting Cura to support the community.

6.4 Governance

Curation is compatible with new modes of online governance [82]. For instance, platforms could set up macro-level norms that are shared by all communities, place global training restrictions on what the curation model could learn to accept, and what content the model should predict to downvote on regardless of the user it is simulating. Or, a platform-wise curation could be applied prior to commmunity-wise curation, and positions as curators could be democratically elected by each community.

Curator selection is essential to make the most of curation: it is decisive for the taste and values of a community. Depending on the curators selected, groups may use Cura to support certain viewpoints or suppress other viewpoints, including those from minorities and historically disadvantaged groups. Existing discussions on who and how to curate content—including social media posts, news feeds and art—have considered professional expertise [62], diversity [26], perspective diversity [19] and other values. Based on our experience in this project, we reflect on four important dimensions that should be taken into consideration during curator selection:

(1) *Democracy.* Is the community selecting all its members as the curators, or only a small set of members as the curators, e.g., only moderators or only the active users? Additionally, democratic values can manifest within the selection of curators: do the community members get to vote or otherwise influence curator choice? [25, 82]
(2) *Expertise.* Are the curators experts in the topic/field of the community? Expert curators can help support high quality community content, but this is in tension with democratic communities.
(3) *Opinion diversity.* Do the curators hold differing viewpoints, or do they have very similar taste? Curators with high opinion diversity can help prevent echo chambers. A community should take into consideration, though, if the curators' preferences have muddled or even strongly conflicting, performing curation may lead to a community without clear norms or purpose.
(4) *Identity diversity.* Are the curators a balanced selection of users with different identities and from different stakeholders groups [31]?

We allow administrators to select curators for simplicity of explanation, but in many communities, they might want curators to be selected through a democratic election. Communities may also explore other possible procedures for selecting curators and different forms of governance [82].

The current Cura system regards every curator equally—the final curation decision is based on the aggregation of one-person one-vote. However, consider editors in media: there are usually multiple editors with one chief editor whose opinion is decisive. Or consider the fan community of a celebrity, with the celebrity themselves also being a member: the voice of the celebrity themselves usually account more than other fan members. In these situations, it may be appropriate to weigh voices unequally. Future research could develop an improved interface that enables administrators to transparently manage their curators and assign different weights to them. The interface could serve as a "staging" area, where administrators can experiment with different weights and balances of curators. They can compare side-by-side feed simulations of different settings to understand the impact of those changes on the community before implementing them live.

6.5 Ethical considerations

In curation, community norms and values are explicitly defined by naming curators. Today's communities still follow the taste of certain users, but that decision is instead implicit, based on the platform design and relative activity level of different community members. Our normative position is that it is better to force this value-laden decision to be made explicitly, where it can be

seen and deliberated, rather than implicitly allowing anti-social behavior to persist under the cover of decentralized community upvoting as in today's platforms. Making this choice explicit does not guarantee fairness or equity; but it gives the community members a better chance to make informed decisions about the communities they invest in.

However, increasing visibility on the curators' votes can cause harm; marginalized community members, possibly selected for their unique perspective, may end up as targets of harassment from those who dislike their curatorial decisions, and displaying their decisions can add fuel to the fire. In practice, we suggest notifying curator candidates about the potential consequences of being selected curators and asking them for affirmative consent before defining them as curators, letting curators' votes be audited, but not necessarily visible publicly so curators have plausible deniability and are not discouraged from expressing their opinions. We also suggest having several curators for contentious groups, so that the curation result is their combined effort and no one becomes a lone target.

Curation centralizes power in the hands of the curators, which is not a value-neutral decision. Centralized power has historically been used to oppress and suppress undesired viewpoints. An approach to mitigating this risk is enabling centralized curation, but requiring that curator selection be subject to democratic election: a technique advocated for in the feminist essay "The Tyranny of Structurelessness" [24].

How do we stop malicious actors from creating communities that curate anti-social or evil content? Platforms can mitigate the risk of this outcome, for example, by placing alarms or restrictions on the kinds of content that a curation model is allowed to learn to upvote. This could be achieved through a hidden test set of objectionable material: if a curator is chosen such that the model would upvote the objectionable material, the platform may refuse to allow that curator, or refuse to allow the model to upvote such material.

Another concern is strategic manipulation of the algorithm, e.g., upvoting toxic content to affect the model's prediction and push specific content to the frontstage. Such behaviors by a small number of people will not change the model's prediction (Figure 8(B)). The system also has some measure of incentive compatibility: if the ill-intentioned users are continuously trying to trick the algorithm, the system will learn that their votes are uncorrelated with the curators' opinions, thus those users will gradually lose their ability to have any impact on the prediction. A more serious threat could involve an individual disguising themselves as a respected community leader and subsequently manipulating voting patterns once selected as a curator by the administrator. Future research could develop automated systems that can detect possible attackers by analyzing inconsistencies in their voting behavior and identifying significant discrepancies from other users' voting histories.

6.5.1 Audits and oversight. Curation models can be audited: an automated audit could regularly test and identify which communities' models are letting hateful content through, and quarantine or deplatform those communities [11, 12]. As mentioned above, platform implementation could also place global training restrictions on what such models could learn to accept, making it difficult for communities to post any hateful content. So, if combined with platform governance, curation models may offer some opportunities here as well.

6.5.2 Filter bubbles and echo chambers. One might worry that curation will harm the content diversity of a community, only presenting content that are from the same perspective, leading to filter bubbles and echo chambers. The community might reinforce extreme and unhealthy behavior, e.g., a beauty community encourages an unhealthy pursuit of beauty through excessive dieting. To mitigate these risks, a few interventions are possible:

(1) Systems might help users visualize the extent to which their communities represent a broad swath of points of view [52].
(2) Communities' administrators might diversify curator selection. A community can have "balancing forces" in curator seats, of the sort that a representational democracy offers, if thought through carefully.
(3) Communities might be connected hierarchically so that no community is an island unto itself: each community's content can percolate up from the metaphorical neighborhood to city to county to country, and vise versa, ensuring that cross-cutting content traverses the network.

However, the same features that invoke fears around echo-chambers also facilitate safe and purposeful spaces; the interventions mentioned above are not solutions for every community. In contrast to the example of a beauty community above, imagine the case of a community for eating disorder recovery, where not only would opposing viewpoints be directly harmful, but content from other communities trickling in (for example, high fashion or modelling) could be as well. These contrasting examples show that there is no single correct decision for all contexts; rather, there are many thoughtful decisions to be made when applying the curation process to different communities.

6.6 Limitations

While we perform technical evaluations and demonstrations of curation, this work does not yet report a field deployment. As a result, we cannot report second-order behavioral effects of curation on a community, nor how its members react to curation. Future work can investigate the emergent effects of curation on community behavior and norms: does it create more norm-adherent behavior? What proportion of content makes it over the curation threshold? Are subjective experiences of curation enjoyable, or frustrated?

Since the present Reddit dataset was sourced from Reddit users who voluntarily made their voting data public, it is possible that our dataset includes a higher proportion of frequent Reddit users, who exhibit above-average rates of voting. Nevertheless, if implemented on a real platform where complete voting data is accessible, the Cura system would have fuller access to all voting behavior that users are willing to share with the system.

The Reddit dataset we use is collected from real users, but not from communities that were originally intended to be curated. So, our results may differ if communities and votes were initially authored with the express purpose of curation. Likewise, Reddit users who voluntarily make their vote data public may differ categorically from other Reddit users, potentially limiting the generalizability of our findings. Since the original dataset is also extremely large and easily exhausts many computational resources, we currently can only make claims about a subset of sampled subreddits. While we restrict our curators to users who have voted at least 5 times, the curation model needs to be finetuned or retrained before being applied to very small communities. Another limitation of our dataset is that it lacks the image and video data in Reddit posts, presenting great challenges for our model to predict curators' votes on posts where images or videos are the main content.

The AI models underlying curation are not perfect. Perfect prediction is impossible, since opinions are never fully predictable and can be subject to social influence. So, it is important that curators can vote manually to override the model's prediction for them, and that the models retrain with curators' votes to improve their performance. However, perfect prediction is also not necessary for curation to succeed, given the inherent noise around social ranking [29].

7 CONCLUSION

We present Cura, a system that enables curation in social media with algorithms and interface. Curation is a metaphor in which the curators of a community decide which content is shared with the larger group, placing trust in that curator's taste to maintain the norms and content quality of the group. Although commonly applied in news media, requiring curators to review every piece of content is too effort-intensive to enable it being instantiated in social media. We overcome this barrier by leveraging a transformer-based deep learning model to predict whether each curator would upvote each post, based on community feedback. We evaluate our approach and demonstrate that our model can accurately estimate curators' opinions and updates quickly as votes from members arrive. We also demonstrate that curation can inflect the same inventory of posts into many different community types, depending on the curators' taste. Curation enables a wide variety of community types ranging from a small group of editors (e.g., newspaper editorial boards), to stakeholder roundtables (e.g., including minoritized groups as curators), to democracies (e.g., giving every user equal voice, instead of just the small number of active users or moderators).

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research. Mitchell Gordon was supported by the Apple Scholars in AI/ML program.

We are grateful to the Chinese Undergraduate Visiting Research Program for providing Wanrong He with the opportunity to undertake this research internship at Stanford University.

We would like to thank the participants of our online experiments, whose feedback greatly contributed to the findings of this research. Their engagement was crucial in shaping the outcomes and enhancing the overall quality of this work.

Additionally, we acknowledge and thank Kris Jeong, Nicole Garcia, and Pauline Arnoud for their collaboration and teamwork throughout this project. They brought fresh ideas and perspectives to the table, broadening the scope and depth of this research.

Finally, we express gratitude to all the individuals, mentors, and colleagues who provided guidance and assistance throughout this research. Their advice, encouragement, and expertise were indispensable in shaping the direction and outcomes of this work.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David Robinson. 2020. Roles for computing in social change. In *FAT**. ACM, 252–260. https://www.microsoft.com/en-us/research/publication/roles-for-computing-in-social-change/

[2] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[3] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics* 4 (2016), 463–476. https://doi.org/10.1162/tacl_a_00111

[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *ArXiv* abs/2004.05150 (2020).

[5] Zane L. Berge and Mauri P. Collins. 2000. Perceptions of e-moderators about their roles and functions in moderating electronic mailing lists. *Distance Education* 21, 1 (Jan. 2000), 81–100. https://doi.org/10.1080/0158791000210106

[6] Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 125–135. https://doi.org/10.18653/v1/2021.insights-1.18

[7] Axel Bruns. 2009. From reader to writer: Citizen journalism as news produsage. In *International handbook of internet research*. Springer, 119–133.

[8] Moira Burke, Cameron Marlow, and Thomas Lento. 2009. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 945–954.

[9] Justin Buss, Hayden Le, and Oliver L. Haimson. 2021. Transgender identity management across social media platforms. *Media, Culture & Society* 44 (2021), 22 – 38.

[10] Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 conference on empirical methods in natural language processing*. 286–295.

[11] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

[12] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[13] Daphne Chang, Erin L Krupka, Eytan Adar, and Alessandro Acquisti. 2016. Engineering information disclosure: Norm shaping designs. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 587–597.

[14] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4743–4754. https://doi.org/10.18653/v1/D19-1481

[15] Andrew Chen. 2021. *The Cold Start Problem: How to Start and Scale Network Effects.* Harper Business, New York.

[16] Zhilong Chen, Jinghua Piao, Xiaochong Lan, Hancheng Cao, Chen Gao, Zhicong Lu, and Yong Li. 2022. Practitioners Versus Users: A Value-Sensitive Evaluation of Current Industrial Recommender System Design. https://doi.org/10.1145/3555646 arXiv:2208.04122 [cs].

[17] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1217–1230.

[18] Robert B Cialdini, Carl A Kallgren, and Raymond R Reno. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*. Vol. 24. Elsevier, 201–234.

[19] Xi Cui and Yu Liu. 2017. How does online news curate linked sources? A content analysis of three online news media. *Journalism* 18 (2017), 852 – 870.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[21] Nicholas Diakopoulos. 2015. Picking the NYT picks: Editorial criteria and automation in the curation of online news comments. *ISOJ Journal* 6, 1 (2015), 147–166.

[22] Brian Dobreski. 2018. Toward a value-analytic approach to information standards. *Proceedings of the Association for Information Science and Technology* 55, 1 (2018), 114–122. https://doi.org/10.1002/pra2.2018.14505501013 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2018.14505501013.

[23] Yan Fan, Chengyu Wang, Peng He, and Yunhua Hu. 2022. Building Multi-turn Query Interpreters for E-commercial Chatbots with Sparse-to-dense Attentive Modeling. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (2022).

[24] Jo Freeman. 1972. The tyranny of structurelessness. *Berkeley Journal of Sociology* (1972), 151–164.

[25] Seth Frey, PM Krafft, and Brian C Keegan. 2019. " This Place Does What It Was Built For" Designing Digital Institutions for Participatory Change. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–31.

[26] Lisa Gaupp. 2021. *12 How to Curate Diversity and Otherness in Global Performance Art.* De Gruyter Open Poland, Warsaw, Poland, 290–321. https://doi.org/doi:10.1515/9788366675308-014

[27] Anna Gausen, Wayne Luk, and Ce Guo. 2022. Using Agent-Based Modelling to Evaluate the Impact of Algorithmic Curation on Social Media. *ACM Journal of Data and Information Quality (JDIQ)* (2022).

[28] Eric Gilbert. 2012. Designing social translucence over social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2731–2740.

[29] Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 803–808.

[30] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[31] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3502004

[32] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[33] James Grimmelmann. 2015. The Virtues of Moderation,. *Yale Journal of Law and Technology* 17 (2015), 42–109.
[34] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[35] Thomas Hofmann. 2004. Latent Semantic Models for Collaborative Filtering. *ACM Trans. Inf. Syst.* 22, 1 (jan 2004), 89–115. https://doi.org/10.1145/963770.963774
[36] Jim Hollan and Scott Stornetta. 1992. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 119–125.
[37] Tiffany W Hsu, Yu Niiya, Mike Thelwall, Michael Ko, Brian Knutson, and Jeanne L Tsai. 2021. Social media users produce more affect that supports cultural values, but are more influenced by affect that violates cultural values. *Journal of Personality and Social Psychology* (2021).
[38] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S Ackerman, and Eric Gilbert. 2021. Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
[39] Harsh Khatter and Anil Kumar Ahlawat. 2020. Analysis of Content Curation Algorithms on Personalized Web Searching. *Social Science Research Network* (2020).
[40] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1152–1156.
[41] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* 1 (2012), 4–2.
[42] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2015).
[43] Yehuda Koren, Steffen Rendle, and Robert Bell. 2022. Advances in Collaborative Filtering. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, New York, NY, 91–142. https://doi.org/10.1007/978-1-0716-2197-4_3
[44] Robert Kraut, Moira Burke, John Riedl, and Paul Resnick. 2010. Dealing with newcomers. *Evidencebased Social Design Mining the Social Sciences to Build Online Communities* 1 (2010), 42.
[45] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.
[46] Francis LF Lee, Michael Che-ming Chan, Hsuan-Ting Chen, Rasmus Nielsen, and Richard Fletcher. 2019. Consumptive News Feed Curation on Social Media as Proactive Personalization: A Study of Six East Asian Markets. *Journalism Studies* 20, 15 (2019), 2277–2292.
[47] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All That's Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 584–595.
[48] G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* 7, 1 (2003), 76–80. https://doi.org/10.1109/MIC.2003.1167344
[49] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.
[50] Lisa Merten. 2021. Block, Hide or Follow—Personal News Curation Practices on Social Media. *Digital Journalism* 9, 8 (2021), 1018–1039. https://doi.org/10.1080/21670811.2020.1829978 arXiv:https://doi.org/10.1080/21670811.2020.1829978
[51] Erwan Moreau, Carl Vogel, and Marguerite Barry. 2019. A Paradigm for Democratizing Artificial Intelligence Research. In *Innovations in Big Data Mining and Embedded Knowledge*, Anna Esposito, Antonietta M. Esposito, and Lakhmi C. Jain (Eds.). Springer International Publishing, Cham, 137–166. https://doi.org/10.1007/978-3-030-15939-9_8
[52] Sean Munson, Stephanie Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of The International AAAI Conference on Web and Social Media*, Vol. 7. 419–428.
[53] Brad Myers, Scott E Hudson, and Randy Pausch. 2000. Past, present, and future of user interface software tools. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7, 1 (2000), 3–28.
[54] Brigitte Naderer, Raffael Heiss, and Jörg Matthes. 2020. The skilled and the interested: How personal curation skills increase or decrease exposure to political information on social media. *Journal of Information Technology & Politics* 17 (2020), 452 – 460.
[55] Chang Sup Park and Barbara K. Kaye. 2019. Mediating Roles of News Curation and News Elaboration in the Relationship between Social Media Use for News and Political Knowledge. *Journal of Broadcasting & Electronic Media* 63 (2019), 455 – 473.

[56] Joon Sung Park, Joseph Seering, and Michael S Bernstein. 2022. Measuring the Prevalence of Anti-Social Behavior in Online Communities. *Proceedings of the ACM on Human-Computer Interaction* (2022).

[57] Michael J. Pazzani. 1999. A Framework for Collaborative, Content-Based and Demographic Filtering. *Artif. Intell. Rev.* 13, 5–6 (dec 1999), 393–408. https://doi.org/10.1023/A:1006544522159

[58] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 557–568.

[59] Paul Resnick and Hal R. Varian. 1997. Recommender systems. *Commun. ACM* 40 (1997), 56–58.

[60] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation.* University of Illinois at Urbana-Champaign.

[61] Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Association for Computational Linguistics, Online, 2182–2194. https://doi.org/10.18653/v1/2021.acl-long.170

[62] Steve Rosenbaum. 2009. Can Curation Save Media. https://www.businessinsider.com/can-curation-save-media-2009-4.

[63] Lee Ross and Richard E Nisbett. 1991. *The person and the situation: Perspectives of social psychology.* Pinter & Martin Publishers.

[64] Costanza-Chock Sasha. 2020. Design Justice. https://mitpress.mit.edu/9780262043458/design-justice/

[65] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376506

[66] Charlotte Schluger, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.

[67] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640.

[68] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. 2018. Applications of social identity theory to research and design in computer-supported cooperative work. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–34.

[69] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.

[70] Oren Sar Shalom, Haggai Roitman, and Pigi Kouki. 2022. Natural Language Processing for Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, New York, NY, 447–483. https://doi.org/10.1007/978-1-0716-2197-4_12

[71] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376783

[72] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. 2022. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. http://arxiv.org/abs/2207.10192 arXiv:2207.10192 [cs].

[73] Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence* 2009 (Oct. 2009), 1–19. https://doi.org/10.1155/2009/421425

[74] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19).* Association for Computing Machinery, New York, NY, USA, 1441–1450. https://doi.org/10.1145/3357384.3357895

[75] Joëlle Swart. 2021. Experiencing Algorithms: How Young People Understand, Feel About, and Engage With Algorithmic News Selection on Social Media. *Social Media + Society* 7, 2 (2021), 20563051211008828. https://doi.org/10.1177/20563051211008828 arXiv:https://doi.org/10.1177/20563051211008828

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[77] Mikko Villi, Johanna Moisander, and Annamma Joy. 2012. Social curation in consumer communities: Consumers as curators of online media content. *ACR North American Advances* (2012).

[78] Yixue Wang and Nicholas Diakopoulos. 2022. Highlighting High-Quality Content as a Moderation Strategy: The Role of New York Times Picks in Comment Quality and Engagement. *Trans. Soc. Comput.* 4, 4, Article 13 (jan 2022), 24 pages. https://doi.org/10.1145/3484245

[79] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[80] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-offs Across Multiple Objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1245–1257. https://doi.org/10.1145/3357236.3395528

[81] Amy X Zhang, Mark S Ackerman, and David R Karger. 2015. Mailing Lists: Why Are They Still Here, What's Wrong With Them, and How Can We Fix Them?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4009–4018.

[82] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: building governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 365–378.

[83] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 194:1–194:23. https://doi.org/10.1145/3274463

[84] Jie Zou, Evangelos Kanoulas, Pengjie Ren, Zhaochun Ren, Aixin Sun, and Cheng Long. 2022. Improving Conversational Recommender Systems via Transformer-based Sequential Modelling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2319–2324. https://doi.org/10.1145/3477495.3531852

A ONLINE EXPERIMENT DETAILS

We list how the fifteen pairs of feeds in our online experiment are generated in Table 9, including the Reddit subreddit being used, and the curator groups that curate the target feed and the incorrect distractor feed, as well as if the distractor feed is generated through typical community upvoting (broadcast) or curation.

Also note that the training example is: r/PoliticalDiscussion curated by members who are also members of r/democrats (correct) vs. r/PoliticalDiscussion curated by members who are also members of r/Republican (incorrect).

Received January 2023; revised April 2023; accepted May 2023

Target Feed	Distractor Feed
r/technology curated by members who are also members of r/programming	r/technology
r/technology curated by members who are also members of r/teenagers	r/technology curated by members who are also members of r/Conservative
r/PoliticalDiscussion curated by members who are also members of r/Conservative	r/PoliticalDiscussion
Super-politics community (sampling 500 posts from each of r/politics, r/Conservative, r/Liberal, r/Republican, r/democrats, and r/PoliticalDiscussion) curated by members who are also members of r/democrats	Super-politics community curated by members who are also members of r/Republican
r/Jokes curated by members who are also members of r/LesbianActually	r/Jokes
r/Jokes curated by members who are also members of r/teenagers	r/Jokes
r/Jokes curated by members who are also members of r/Conservative	r/Jokes
r/teenagers curated by members who are also members of r/gaming	r/teenagers
r/teenagers curated by members who are also members of r/travel	r/teenagers curated by members who are also members of r/punk
r/worldnews curated by members who are also members of r/Liberal	r/worldnews
r/worldnews curated by members who are also members of r/india	r/worldnews curated by members who are also members of r/france
r/gaming curated by members who are also members of r/teenagers	r/gaming
r/gaming curated by members who are also members of r/LesbianActually	r/gaming curated by members who are also members of r/scifi
r/music curated by members who are also members of r/Christianity	r/music curated by members who are also members of r/scifi
Super-science community (sampling 500 posts from each of r/science, r/ScienceFacts, r/technology, and r/shittyaskscience) curated by members who are also members of r/programming	Super-science community curated by members who are also members of r/Jokes

Table 9. Generation details of the fifteen pairs of feeds.