The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design

Hong Shen
hongs@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Leijie Wang
wanglj17@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Wesley H. Deng
hanwend@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Ciell Brusse
ciell.wikipedia@gmail.com
Dutch Wikipedia
Gld, Netherlands

Ronald Velgersdijk
wikipedia@ronaldvelgersdijk.nl
Dutch Wikipedia
Gld, Netherlands

Haiyi Zhu
haiyiz@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

There have been increasing calls for centering impacted communities – both online and offline – in the design of the AI systems that will be deployed in their communities. However, the complicated nature of a community's goals and needs, as well as the complexity of AI's development procedures, outputs, and potential impacts, often prevents effective participation. In this paper, we present the *Model Card Authoring Toolkit*, a toolkit that supports community members to understand, navigate and negotiate a spectrum of machine learning models via deliberation and pick the ones that best align with their collective values. Through a series of workshops, we conduct an empirical investigation of the initial effectiveness of our approach in two online communities – English and Dutch Wikipedia, and document how our participants collectively set the threshold for a machine learning based quality prediction system used in their communities' content moderation applications. Our results suggest that the use of the *Model Card Authoring Toolkit* helps improve the understanding of the trade-offs across multiple community goals on AI design, engage community members to discuss and negotiate the trade-offs, and facilitate collective and informed decision-making in their own community contexts. Finally, we discuss the challenges for a community-centered, deliberation-driven approach for AI design as well as potential design implications.

ACM Reference Format:
Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533110

1 INTRODUCTION

Algorithmic systems powered by machine learning techniques have increasingly exercised power over a wide variety of communities, both online and offline. Wikipedia, for example, uses algorithmic tools to assess the quality of edits in articles and to take appropriate actions based on this editing quality, such as reverting problematic edits [30]. Algorithmic tools have also been deployed in many offline, local communities to help, for example, child protection agencies screen referral calls [21], or school districts redesign their bus driving routes [10]. Those AI tools and systems that appear to provide efficient *technical* solutions, however, can fail in practice, often due to disconnections between the designs of these systems and real-world community stakeholders' realities, context, and constraints, which is both likely to undermine practical initiatives and to cause significant harm to already disadvantaged social groups [20, 28, 33, 64].

Recently, there have been increasing calls for centering impacted communities in the design of the AI systems that will be deployed in their communities (e.g., [14, 37, 40, 43, 53, 67, 72]). As Shah [55] has argued, achieving "social license" from the broader community is critical to the ability of technologies to have a positive social impact.

However, the diverse and sometimes contradictory nature of a community's goals and needs [60], and the complexity of AI's development procedure, outputs, and potential impacts [16, 44, 45, 68] have prevented community members from effective participation in system design. In particular, there are often inherent *trade-offs* in implementing multiple community goals in the design of AI systems. Optimizing one goal can lead to poor performance on others. For example, there is a well-documented trade-off between false positives and false negatives: reducing false positives can increase false negatives and vice versa. Furthermore, research has also shown a trade-off between prediction accuracy and fairness [38, 39]. The complex model trade-offs, coupled with the diverse values and goals embedded in any given communities, make it challenging for community members to understand, navigate, and negotiate the trade-offs associated with different design choices, and make collective decisions about what is best outcomes for their communities.

In this paper, we present the *Model Card Authoring Toolkit*, a toolkit that supports community members to understand, navigate and negotiate a spectrum of machine learning models via deliberation and pick the ones that best align with their collective values. Instead of algorithmically aggregating individual's preferences for AI design (e.g., [36, 43]), we argue for the importance of collective

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533110

decision-making via deliberation, mediated through technical innovations [56, 58, 72]. The Model Card Authoring Toolkit encompasses three different artifacts – the Authoring Tool, the Comparison Table, and the Deliberation Protocol, to scaffold the deliberation process. The *Authoring Tool* allows community members to navigate a spectrum of machine learning models and choose a model that aligns with their values. The *Comparison Table* concisely summarizes the key performance metrics of different models and explains their impacts on different stakeholder groups. The *Deliberation Protocol*, which includes an onboarding session, a discussion session, and a two-step writing session, helps structure the discussion process and facilitate collective decision-making.

In this paper, we document an early use of our approach in the Wikipedia community. Through a series of workshops, we conduct an empirical investigation of the initial effectiveness of our approach in two different Wikipedia communities – English Wikipedia (enwiki) and Dutch Wikipedia (nlwiki), and document how our participants try to collectively set the threshold for a machine learning based quality prediction system used in their communities' content moderation applications. Our results suggest that the use of the Model Card Authoring Toolkit helps improve the understanding of the trade-offs across multiple community goals on AI design, engage community stakeholders to discuss and negotiate the trade-offs, and facilitate collective and informed decision-making in their own community contexts. Finally, we discuss the challenges for a community-based, deliberation-driven approach for AI design as well as potential design implications.

Our contributions are three-fold:

- First, we introduce a novel deliberation driven toolkit – the Model Card Authoring Toolkit – to help community members understand, navigate and negotiate a spectrum of models via deliberation and try to pick the ones that best align with their collective values;
- Second, we document an early use of our approach in two different Wikipedia communities (English and Dutch) and conduct an empirical investigation of the initial effectiveness of our approach;
- Third, we discuss the challenges for a community-centered, deliberation-driven approach for algorithm design as well as potential design opportunities.

2 RELATED WORK

In this section, we outline relevant work in two areas. First, we present an overview of participatory approaches to AI and machine learning, a field that has received increasing attention in FAccT and describe how our work contributes to this emerging line of research. Next, we survey the existing techniques, toolkits, and systems related to issues of FATE in the fair ML literature, and describe how our work is positioned in this space.

2.1 Participatory Approaches to AI and Machine Learning

Recently, recognizing the limitations of merely offering *technical* perspectives to system design, works in HCI and FAccT have started to incorporate a wide range of stakeholders into the design process, often under the umbrella term of "participatory approaches to AI and Machine Learning" [41]. For example, past work has built systems that ensure users have the potential to adapt their predictions or recommendations [22, 63] and actively incorporate their feedback into system design [69]. They have developed methods and tools for informing and engaging users in model design, development and deployment [47, 51] as well as for them to analyze and contest the outcomes of ML systems [42]. They have also proposed methods and frameworks for developers, researchers and everyday users to detect, understand and interrogate problematic machine behaviors [2, 15, 57]. Despite their significant contribution, communities that are mostly impacted by decision-making algorithms have often not been sufficiently centered in the process [34, 59]. As a result, there have been increasing calls for re-centering impacted communities – stakeholders who are at the receiving end of the AI decisions and more likely to be significantly harmed – in the design of the AI systems that will be deployed in their communities (e.g., [14, 37, 40, 43, 53, 67, 72]).

Our work answers this call and complements previous research in participatory approaches to AI and Machine Learning by re-centering community values and involving real-world community members in the design process of decision-making algorithms. In particular, we focus on *trade-off deliberation* to facilitate collective community discussion and decision-making in system design. In this project, we developed and evaluated a novel deliberation-driven toolkit – the Model Card Authoring Toolkit – to help community members understand, navigate and negotiate different model trade-offs across multiple community goals on AI design. We evaluated the initial effectiveness of our toolkit in two different Wikipedia communities and discussed challenges and limitations.

2.2 Techniques, toolkits and systems for FATE in AI and Machine Learning

Recently, the fair ML community has also undertaken significant efforts to develop techniques, toolkits, and systems to assist ML development teams in assessing and addressing machine biases (e.g., see [3, 19, 26, 33, 52]). In response to increased concerns about harmful algorithmic behaviors in AI systems (e.g., see [4, 7, 23]), this body of work often aims at addressing FATE – Fairness, Accountability, Transparency and Ethics – related concerns in AI and ML.

On the one hand, sophisticated tools and systems have been developed to help developers and industrial practitioners understand, assess and mitigate machine biases. For instance, Microsoft developed the Fairlearn toolkit [12] and IBM developed the AI Fairness 360 toolkit [8] to help industrial practitioners assess and mitigate potential algorithmic biases presented in their ML models. In a similar vein, Google has developed an open-source tool called "What-if" that helps developers without formal ML training explore the effects of different fairness metrics [66]. On the other hand, a different line of research has explored ways to enhance transparency and center social impacts in model and data reporting. For example, using multidisciplinary documentation techniques, the Model Card approach [47] uses short documents to provide benchmarked evaluation, intended use context, details of performance evaluation procedures, and other relevant information of a trained ML model. The Datasheets approach [27] focuses more on

the datasets, clarifying the characteristics of the data feeding into the ML model. Taking one step further, Value Cards [56] uses a set of cards to inform computer science students and future practitioners the social impacts of different ML models.

Here, we take a complementary angle, looking at designing FATE related toolkits to support community members in the design of their AI systems. Inspired by Value Sensitive Algorithm Design [72], a design process seeking to better incorporate stakeholder values in the creation of algorithmic systems, we introduce the Model Card Authoring Toolkit – a toolkit to help community members understand, navigate and negotiate a spectrum of diverse and oftentimes competing machine learning models via deliberation and try to pick the ones that best align with their collective values.

3 DESIGN OF THE MODEL CARD AUTHORING TOOLKIT

Below, we describe the general design, rationale and objectives of the Model Card Authoring Toolkit. Inspired by the Model Card [47] approach that uses documentation techniques to offer information of a trained ML model to facilitate ethical consideration for practitioners, we envision the Model Card Authoring Toolkit as a versatile toolkit that can be used to provide concise summaries of the model performances, connect them to community goals in AI design, and facilitate deliberation and collective decision-making about model choices for real-world community members.

3.1 General Design Rationale of the Toolkit

Deliberation is at the core of the Model Card Authoring Toolkit. Instead of viewing human values in AI systems as individual dilemmas that can be calculated as aggregations of individual preferences (e.g.,[36, 43, 61]), here, we foreground the importance of collective decision making via deliberation [53, 56, 58, 72]. As Robertson and Salehi argued [53], solely aggregating individual preferences is insufficient to support truly participatory, democratic governance of algorithmic systems. Deliberation refers to a particular sort of discussion — "one that involves the careful and serious weighing of reasons for and against some proposition" [24] — and is an approach to politics in which members from the general public are involved in collective decision-making through the exchange of ideas and perspectives via rational discourse [1, 17]. The value of deliberation before making a decision includes (1) sharing views on a subject that voting or preference aggregation does not allow; (2) considering a wider range of options or new alternatives; (3) encouraging more public-spirited proposals in contrast to those motivated by self-interest; (4) increasing the legitimacy of the ultimate decisions; and (5) improving the moral or intellectual qualities of the participants [1, 24, 65].

Taken as a group, the Model Card Authoring Toolkit is designed specifically to achieve the following **three objectives** to support community-centered algorithm design: (1) it aims at helping community members understand the trade-offs across multiple community goals on AI design; (2) it aims at helping community members discuss and negotiate the trade-offs across a spectrum of models; (3) it aims at helping community members make collective and informed decisions about model choices within their specific community contexts.

To achieve these goals, in the design of the Model Card Authoring Toolkit, we used **three different artifacts** – the Authoring Tool, the Comparison Table, and the Deliberation Protocol, to scaffold the deliberation process. Specifically,

- The *Authoring Tool* offers an interactive interface, describing a range of machine learning model by showing their important performance metrics that reflect the trade-offs across multiple community goals. It allows community members to navigate a spectrum of models and select one that aligns with their goals and values.
- The *Comparison Table* concisely summarizes the key performance metrics of different models and explains their impacts on different stakeholder groups. It allows community members to systematically and comparatively examine a range of models and scaffold the trade-off discussion.
- The *Deliberation Protocol*, which includes an onboarding session, a discussion session, and a two-step writing session, helps structure the discussion process and facilitate collective decision-making.

In the following sections, we will discuss how the design rationale of the toolkit is being manifested in the study context of the Wikipedia community.

3.2 Study Context: ORES - Machine Learning Service in Wikipedia

As Wikipedia's content and readership continue to grow, there is a clear need to help reduce the sheer volume of human labor necessary to maintain the online encyclopedia. Objective Revision Evaluation System (ORES) – a quality prediction system – has been used to assess the quality of edits and take corrective actions. Maintained by the Scoring Platform Team at the Wikimedia Foundation,[1] ORES supports a variety of content moderation tools, such as Recent Changes [2] and Huggle tool,[3] that allow members to review and revert recent revisions in Wikipedia articles [46].

ORES is a machine-learning based algorithmic system trained with human-labelled Wikipedia revisions and outputs a continuous quality score between 0 and 1 for each edit [29]. The ORES API takes in an edit revision ID and generates a "damaging" and a "good-faith" score, which indicate the likelihoods that an edit is damaging or malicious. To successfully deploy these tools, it is important to select the appropriate threshold so that edits with quality scores above the threshold will be identified as damaging. However, there are *trade-offs* associated with threshold selection. For instance, a low threshold ensures that the majority of damage will be caught, but at the expense of needing to review more edits. In contrast, a high threshold minimizes the harm of automatically reverting good edits but may allow a large number of damaging edits through.[4] In most cases, the exact number of the threshold is selected by Wikipedia administrators and is not directly accessible to most Wikipedia members. However, as [29] noted as the development team behind ORES, as a socio-technical system situated in a large online community, the design, development and

[1] https://mediawiki.org/wiki/Wikimedia_Scoring_Platform_team
[2] https://en.wikipedia.org/wiki/Help:Recent_changes
[3] https://en.wikipedia.org/wiki/Wikipedia:Huggle
[4] https://www.mediawiki.org/wiki/ORES/Thresholds

deployment of ORES should also actively involve a wide range of Wikipedia community members. We therefore chose ORES as our study context to design and evaluate our toolkit as a "proof-of-concept" study to understand the challenges and opportunities of community-centered algorithmic design.

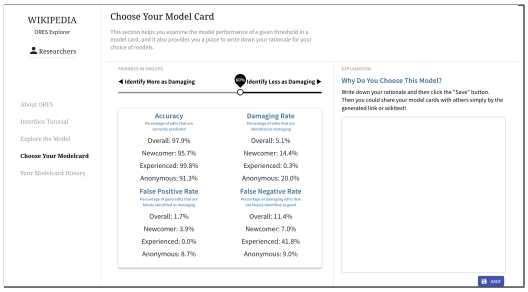
3.3 How the Model Card Authoring Toolkit is being used in the Wikipedia Context

In our case study with ORES, we developed the toolkit based on the ORES explorer [70], which provides users with a suite of visualization interfaces that explains the trade-offs in the ORES system, based on methods developed in [71]. We feed the toolkit with real datasets from the Wikipedia communities and the spectrum of models developed in [70].

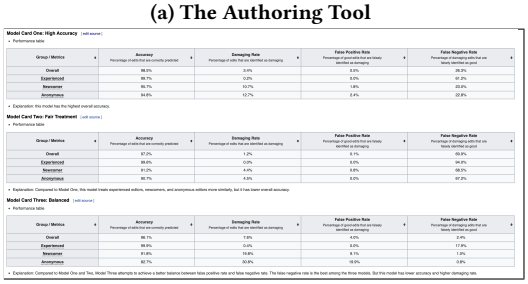
As noted in [60], when developing a machine-learning-based quality predictive tool to help moderate content, the Wikipedia community may have multiple goals: (1) catching all the potentially low-quality edits, (2) avoiding falsely flagging high-quality edits as low-quality, and (3) treating different editor groups fairly and equally. However, as with many other decision-making algorithms deployed in real-world communities, there are often inherent *trade-offs* in implementing multiple design goals in the algorithm. Optimization for multiple criteria is challenging: Optimizing one criterion often leads to poor performance in other criterion. For example, the value "reducing the effort of community maintenance" corresponds to criteria of maximizing overall accuracy and minimizing false positives (i.e., catching all the potential damaging edits); while "encouraging positive engagement with diverse editor groups" corresponds to minimizing false negatives (i.e., not falsely labeling good edits as damaging) and treating different editor groups fairly. However, there is a well-documented trade-off between false positives and false negatives; reducing false positives can increase false negatives and vice versa. Furthermore, machine learning research has shown a trade-off between prediction accuracy and fairness (e.g., [38, 39]). Specifically, improving fairness – such as minimizing differences in false-positive rates between different editor groups – can lead to a decrease in overall prediction accuracy.

For the purpose of this study, we focus on two sets of trade-offs between four different sets of system criteria and their respective community values [60, 70]: the trade-offs between minimizing false positives (i.e., catching all the potential damaging edits) and minimizing false negatives (i.e., not falsely labeling good edits as damaging), and the trade-offs between overall accuracy and fairness (i.e., treating different editor groups equally). Following [70], we also illustrate the model impacts on three different editor groups: newcomers, experienced and anonymous editors. The ultimate goal is to enhance the Wikipedia community's agency in AI design and improve community capability to choose the model that aligns with their collective goals and needs.

3.3.1 The Authoring Tool. The first artifact in the Model Card Authoring toolkit is the Authoring Tool (see Figure 1, Left). It allows participants to create their own model card for a given machine learning model in two panels: The model panel and the social panel. The model panel on the left offers an interactive interface for a set of machine learning models by showing its performance metrics in aggregate and across different social groups to capture and scaffold the potential trade-offs. Participants can use the model panel to navigate a spectrum of models via the selection of different thresholds and pick one that best aligns with their values. After picking the model, they will use the social panel on the right to elaborate on their reasons behind the choice. The interface will then combining the model panel and social panel to generate a Model Card. The aim of the Authoring Tool is to scaffold the potential trade-offs in a machine learning model and allow participants to map their social values to model trade-offs. There are often inherent trade-offs in implementing multiple goals in AI design (e.g., [38, 39]). Moreover, different community groups can have a wide range of diverse goals and values associated with an algorithm. The Authoring Tool helps capture and concretize those implicit model trade-offs and map them back to the diverse goals and values of community members.



(a) The Authoring Tool



(b) The Comparison Table

Figure 1: The Model Card Authoring Toolkit uses three different artifacts to scaffold the deliberation process, including the Authoring Tool, the Comparison Table and the Deliberation Protocol. Figure 1 shows the first two artifacts designed for our case study: the Authoring Tool (Left) and the Comparison Table (Right).

3.3.2 The Comparison Table. The second artifact in the toolkit is the Comparison Table (see Figure 1, Right). Based on individual model cards created by participants using the Authoring Tool, the Comparison Table further scaffolds the resulting set of diverse machine learning models, concisely summarize their key performance metrics, and explain their impacts on different community groups. Inspired by Mitchell et al. [47], the purpose of the Comparison Table is twofold. First, it offers a compact trade-off overview across different machine learning models in an algorithmic system to facilitate systematic examination. Second, it illustrates the social impacts of the model trade-offs on different social groups to allow comparative

discussion. We hope that participants would take the Comparison Table as a starting point to develop systematic understanding of the trade-offs across a spectrum of models and be ready to engage in the deliberation.

3.3.3 The Deliberation Protocol. The third artifact in the toolkit is the Deliberation Protocol. The Wikipedia community has already developed a number of formal and informal deliberation mechanisms for resolving conflicts in content-related disputes and user conduct disputes [35], such as Third Opinion, Dispute Resolution Noticeboard, Mediation Committee, Request for Mediation, and Request for Comments. We reviewed these existing consensus-building mechanisms in our research group iteratively and designed the final protocol to facilitate community-centered deliberation on ORES. Our finalized protocol includes the following three steps: An onboarding session, a discussion session, and a two-step writing session. (1) Participants will first have *an onboarding session*, during which they will be briefed about ORES, the machine learning service used in their Wikipedia communities, including the background, the purpose and the functionality of the system. (2) In the *discussion session*, participants will be encouraged to discuss with each other on which model is producing the best outcomes and would recommend for the community to use. Participants will be encouraged to discuss the following questions: What is the definition of good outcomes in their community? What are the pros and cons of different models? Which model would the group collectively choose? (3) After the discussion, participants will be asked to perform a *two-step writing task*. They will first be asked to summarize and write down the high-level principles that the ORES developers (or any future AI system designers) should consider to better benefit the Wikipedia community. They will then be asked to write a group proposal about their collectively chosen models. Consensus is not required at this point: If they agree on a model, they can include the model and their rationale. If they are not able to agree upon a model, they can propose up to three models, and provide their reasons.

3.4 Limited Scope of the Case Study

The goal of our study is not to capture all model trade-offs in content moderation algorithms used in Wikipedia, but to study the initial effectiveness of the Model Card Authoring Toolkit as a community-centered, deliberation-driven toolkit in a real-world online community. Therefore, we choose to focus on a restrictive set of model performance metrics, stakeholder groups and editor communities, based on previous work developed in this domain [60, 70].

4 A CASE STUDY OF USING THE MODEL CARD AUTHORING TOOL IN PRACTICE

Below, we describe a case study of using the Model Card Authoring Toolkit in two different Wikipedia communities – English and Dutch Wikipedia – to facilitate community members' understanding, navigation and negotiation of different model trade-offs and collectively set the threshold for ORES. We conducted four online workshops – two with English and two with Dutch using our toolkit.

4.1 Community Context: English and Dutch Wikipedia

ORES has been designed and deployed in 34 different language versions; each language version is an independent editor community [6, 32], with their own editor base, rules, policies, and practices. These 34 communities include large communities like English Wikipedia (enwiki), German Wikipedia (dewiki), to relatively smaller communities such as Dutch Wikipedia (nlwiki). Our study focuses on two different communities – English Wikipedia (enwiki) and Dutch Wikipedia (nlwiki). We chose these two communities for our case study because they differ in size, editor groups, and content moderation policies so that we can observe the effectiveness of the toolkit in different types of online communities (see Table 1 for a brief summary of their different community characteristics).

4.2 Method

4.2.1 Study Procedure and Process. We conducted our workshops via online conferencing platform Zoom. In each workshop, participants were asked to complete a 5-step activity using the Toolkit: (1) Complete a brief pre-survey to understand their role on Wikipedia, prior experience with ORES, and their perception towards ORES; (2) Use the Authoring Tool to choose one model, describe their values behind the choice and generate their individual Model Card; (3) Use the Comparison Table to discuss in groups on which model they would recommend for the community; (4) Write a group proposal about their collectively chosen models; (5) Complete a post-survey. Participants were given the Authoring Tool in step 2, Comparison Table and in step 3, and guided by the Deliberation Protocol throughout the entire workshop. Although ORES can be used for a wide range of content moderation apps in Wikipedia, we asked our participants to focus on selecting models for ORES in the context of Recent Changes,[5] which is one of the most popular pages for Wikipedia members to review and revert recent revisions (See figure 2 for study design).

4.2.2 Data Collection and Demographics. We conducted two deliberation workshops respectively with Dutch and English Wikipedians. We used snowball sampling [11] to recruit both English and Dutch members by posting recruiting messages on individual Wikipedia mailing lists and by reaching out via personal networks. For Dutch Wikipedia, we also recruited using the mailing list and news letters of Wikimedia Netherlands.[6] For English Wikipedia, we recruited additional participants by posting a discussion thread on Village Pump, a discussion forum to discuss the technical issues, policies, and operations of Wikipedia.[7] Each participants were offered $30 (28.88 Euro) Amazon gift cards as a sign of thank you for their volunteer efforts. Some participants from the Dutch community politely declined the gift cards to conform with their cultural values. Each workshop was 120-180 minutes long. All participants were briefed about the study before attending the workshop and explicit consent was obtained. All workshop sessions were audio recorded, in accordance with the Institutional Review Boards of the university.

[5] https://en.wikipedia.org/wiki/Help:Recent_changes
[6] https://meta.wikimedia.org/wiki/Wikimedia_Nederland/The_Netherlands_and_the_world
[7] https://en.wikipedia.org/wiki/Wikipedia:Village_pump

	Dutch Wikipedia	English Wikipedia
Wikipedia Size	6000 edits made every day	160,000 edits made everyday
Community Size	1250-1500 active editors	40k active editors
Editor Groups	Consists mainly of Dutch, highly educated men aged over 40. Only 10% of the editors are women and 3% of the editors have a non-Western migration background	Consists of editors from different countries: editors from top five countries (i.e., US, UK, Canada, India, and Australia) account for 70% of all editors. Only 16% of the editors are women
	Most editors are located in the same time zone, actively communicating with each other via both online and in-person events	Editors are widely located in different time zones all over the world with little possibilities to meet up in real life
Content Moderation	All anonymous edits are checked by humans. About 8-10 moderators/patrollers manually and structurally check all anonymously edits on a daily basis	Checked by patrolling bots and real humans but no structured patrolling system of all anonymous edits

Table 1: A brief summary of different characteristics of English and Dutch Wikipedia in regard with community size, editor groups and content moderation structures. (Table compiled by the authors, with information collected from both English and Dutch Wikipedia as well as the Wikimedia Foundation).

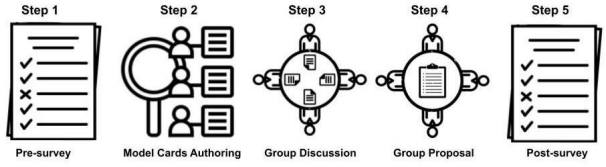


Figure 2: Five Steps of Study Design. In each workshop, participants were asked to complete a 5-step activity: (1) Complete a brief pre-survey; (2) Use the Authoring Tool to generate their individual model card; (3) Use the Comparison Table to discuss in groups on which model they would recommend; (4) Write a group proposal about their collectively chosen models; (5) Complete a post-survey.

In total, there were 15 participants, six from English Wikipedia and nine from Dutch Wikipedia. One participant chose to opt out the surveys, leaving us with 14 survey responses. Participants assumed various and multiple roles in Wikipedia, including editors, content moderators/patrollers, developers, administrators, and mentors. Specifically, *editors* are directly affected by the moderation decisions of ORES-supported systems. *Content moderators/patrollers* are the frequent users of ORES-supported moderation systems. *Developers* build software and tools based on the ORES prediction service. *Administrators* are granted advanced access to perform certain tasks (e.g., blocking users). Mentors support new editors via answering questions. Table 2 contains voluntarily disclosed participant information.

4.2.3 Data Analysis. For data analysis, we used qualitative measures to assess the initial effectiveness of our approach. Due to limited sample size, results from pre- and post-surveys were only compared to offer descriptive statistics. The qualitative datasets used in this study include discussion transcripts, group proposals, and the open-ended questions in the surveys. All the English group discussions were audio-recorded and transcribed using otter.ai [49]. All the Dutch workshop discussions were audio-recorded, transcribed and translated via a third-party service, and validated by the native Dutch authors of this paper. In this case study, our goal is to examine the effectiveness of the toolkit with the three objectives listed in 3.1. To achieve this goal, we used both inductive and deductive coding [13]. Our analysis started from inductive coding to extract codes that show evidence of the deliberation effects. Two authors first read the dataset separately and held weekly discussions. They then coded the first 10% individually and met to reach agreement on the codes. One author then coded the remaining using the codes developed and discussed with the second one iteratively to resolve emergent issues (e.g., ambiguities or new codes). We developed a codebook based on the analysis, summarizing 6 main codes and 33 subcodes. After inductive coding, we conducted a deductive coding in our research group, applying the codebook on the entire dataset to present evidence of deliberation along with the three objectives. Our approach builds upon and differs from existing grounded theory [18] and thematic analysis [13] in that the goal of the inductive coding is not to summarize emerging themes, rather to develop a codebook that serves as a vehicle for the subsequent deductive coding to undercover evidence of the deliberation along with the three objectives.

4.2.4 Research Stance. Recent research in HCI, ML, and FAccT research communities highlighted the potential harms researchers might cause when carelessly conducting community-based research [31, 34, 50]. Specifically, the Wikipedia community has specific rules and norms regarding related research projects. For example,

ID	Gender	Experience	Roles
D1	M	1+ years	Editor & Moderator
D2	Nonbinary	18 years	Editor
D3	M	9 years	Editor
D4	-	-	-
D5	M	17 years	Editor & Moderator
D6	M	7 years	Mentor
D7	F	8 years	Mentor
D8	M	1 year	Editor & Administrator
D9	M	16 years	Editor & Mentor
E1	F	2 years	Mentor & Administrator
E2	M	0.5 year	Administrator
E3	F	3 years	Editor
E4	M	4 years	Mentor & Administrator
E5	F	11 years	Editor & Moderator
E6	M	14+ years	Editor & Developer

Table 2: Participant Summary (N=15). A single dash indicates that the participant chose to participate in the workshop but not the surveys, leaving us with 14 survey responses. Our participants assumed various and multiple roles in Wikipedia, including editors, content moderators/patrollers, mentors, administrators, and developers. We use "E" to represent participants from the English Wikipedia and "D" to identify participants from the Dutch Wikipedia.

"Wikipedia is not a laboratory".[8] Before starting this project, we consulted multiple experienced Wikipedia editors, algorithm developers in the Wikimedia community, and Social Computing scholars studying Wikipedia to ensure that we respected and followed the community norms when recruiting and working with editors. During the research progress, we were transparent about our research goals to the Dutch and English community members and actively built rapport with them. We also created a research page to disclose our research purposes, progress, results and encourage open discussion on Meta-Wiki,[9] the site for the Wikimedia Foundation to coordinate and document relevant research projects.

4.3 Results

In general, our workshop results suggest that the use of the Model Card Authoring toolkit is promising in achieving the **three objectives** listed in 3.1. It helps (1) improve the understanding of the trade-offs across multiple community goals on AI design, (2) engage community stakeholders to discuss and negotiate the trade-offs, and (3) facilitate collective and informed decision-making in their own community contexts. Below, we report our findings through the qualitative analysis aggregating the deliberation transcripts, group proposals, and open-ended questions in the surveys as well as descriptive statistics comparing the results from the pre- and post-surveys. We use "E" to represent participants from the English Wikipedia and "D" to identify participants from the Dutch Wikipedia.

[8]https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not
[9]https://meta.wikimedia.org/wiki/Main_Page

4.3.1 Improving understanding of the trade-offs across multiple community goals on AI design. Overall, we found that the use of the toolkit improved participants' understanding of the model trade-offs and enabled them to connect those trade-offs to a wide range of community goals and values.

First, a clear pattern has emerged that during the deliberation, **participants often drew from their lived experiences in the Wikipedia communities to contextualize model trade-offs**, which helped them quickly connect *technical* trade-offs with the *social* goals of the AI system they have in mind. For example, in the first Dutch workshop, after using the Authoring Tool, D4 talked about his experience as a long-time content moderator in the community. Since he knows the majority of the editors in this relatively smaller community, if new ones come in, he can patrol their edits immediately so his goal with ORES is to have lower Positive Rate: *"You know, I've been working in this community for a long time and I'm relatively familiar with the editors here. I would like the system to mark less edits as damaging – low positives, right? – as I can catch bad edits pretty quickly."*

Similarly, during the second English workshop, E6 advocated for "minimizing false positive rate" since as a long-time Wikipedia editor, *"I only review Recent Changes from time to time because I want to be disturbed by moderation work as little as possible. I think ORES should only bother me with content that highly relevant to the topic ... (I'd prefer) having a minimized false positive rate so there will be less judgement calls to make."*

Based on her own editing experiences, E5 disagreed with E6 and advocated for a higher false positive rate, she said, *"For me, when I review and edit, I always go to the Recent Changes and then decide for myself. So in that case, what I would want as an editor is for the system (ORES) to show me a lot, a lot of live edits, and I would go into each of them in detail, decide if it's a good edit or a damaging edit."*

Second, we also noticed that **participants were able to not only draw from their own experience when analyzing the trade-offs, but also consider *other stakeholders'* needs**, suggesting an improved understanding across multiple community goals on AI design. For example, as a content moderator with 11 years in Wikipedia, E5 advocated for building a "safety net" for new moderators who "are more likely to make errors in judgment" when reviewing the edits. Though E5 preferred to receive more flagged posts herself, she suggested minimizing false positive rate for new moderators so they were not overwhelmed by the large amount of decisions to make.

Similarly, in one of the Dutch workshops, D4 realized that there exist a diverse set of goals and values in his community regarding the design of ORES. He said: *"Look, if someone says: "I have nothing to do in the next six hours, it can take a while and I'll look at it in great detail", then I'd choose D1's model, but if I only had one hour, I'd choose D2' model and if I had even less then I'd choose D3's model. So different people will have different positives or negatives."*

In both the pre- and post-survey, we also asked our participants about their perceived understanding of, trust toward, satisfaction and control of the ORES system. Our data has shown that after the workshop, 14 participants' perceived understanding of the ORES system increases from 5.36 (SD=3.10) to 7.64 (SD=1.74) on a ten-point scale, which are consistent with our findings from qualitative analysis. In addition, there is a slight increase in trust toward the

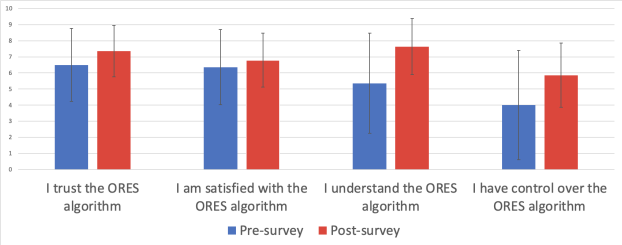


Figure 3: The Changes of Participants' Perception of the ORES System (N=14). The y-axis represents the average rate participants give to each statement on the x-axis out of a 1-10 scale, with the error bar indicating the standard deviation. Figure 3 shows that our toolkit increases participants' understanding of, trust towards, satisfaction and control about the ORES system.

ORES system among the participants from 6.5 (SD = 2.28) to 7.36 (SD = 1.60). However, due to the limited sample size of our qualitative study, we cannot draw statistical inferences.

4.3.2 Facilitating the discussion and negotiation of the trade-offs. We also noticed that the use of the toolkit has helped participants to discuss and negotiate the model trade-offs across multiple design goals.

First, we found out that after engaging in the deliberation, **participants quickly realized that what they were facing was a "dilemma," that is, there was no single perfect solution to the problem, rather, they need to strike a balance between multiple community goals in the design of the ORES system.** For example, D1 sees this as a dilemma between catching as much vandalism as possible and the limited time he has as a content moderator: *"I personally find it important that there are relatively little false negatives without having a really long list of things I need to do. I have limited time, but on the other hand I also don't want to let vandalism go unnoticed."*

D2 understands this as a need to create an "orange light" as a middle ground, in addition to the "green" or "red" light: *"This is why I vote for a stoplight with 3 lights instead of 2, and then use both the low percentage that D3 chooses and the high percentage that D4 chooses all the while making an orange light as the middle field."* D7 summarized the process as "a question of balance": *"For me it was more a question of balance, like water with wine."*

Similarly, in the English workshop, E4 also concluded: *"There isn't a solution, just a trade-off between different positives and negatives ... We're struggling to find a default because what a default reviewer wants to see varies a lot... a default review might be the one that doesn't pick any specific setting."*

Second, we observed that **participants also started to change their priorities and model choices after listening to each other's perspective, suggesting an active process of perspective taking.** For example, in the second Dutch workshop, D7, an editor with 8-year experience changed his choice after engaging with other participants in the workshop: *"Yes, my initial choice was based on the importance of accuracy and decreasing workload, but now that I hear from you that the administrators don't prefer this, I would rather choose model 2."*

After the first English workshop, E1 also appreciated the opportunity to participate in the workshop and thought *"it's been very useful for us to reflect on various perspectives"*. E5 commented that she: *"didn't know the other editors personally, but the discussion with them was fruitful and it helped me refine my perspectives. The discussion with the group has impacted my final choices"*.

In both our pre- and post-survey, we asked our participants about their priorities in maintaining content quality on Wikipedia. We noticed that in the pre-survey, 9 out of 14 tend to consider that "maximizing overall accuracy of the system" as of higher importance and the editor groups in general have divergent goals toward model design. After the deliberation workshop, the participants started to change their priorities. In particular, they were less likely to prioritize accuracy as the most important goal for content moderation, with only 5 out of 14. In addition, they have also narrowed down their listed priorities with the average number of selected priorities decreasing from 4.89 to 3.67, which suggested a potential convergence of different design goals after our workshops.

4.3.3 Enabling collective and informed decisions within specific community contexts. We observed that our toolkit has helped participants to make collective decisions of the ORES design within their specific community context.

In our deliberation workshop, we asked our participants to perform a two-step group writing session after discussion. We first asked them to write a list of high-level principles that the future AI designers in the Wikimedia community should consider; we then asked them to write a group proposal about their collectively chosen model, if they have reached agreement after discussion. We noticed that this task design has helped our **participants to summarize their values and consolidate their diverse goals on the design of the AI system.**

For example, participants from our Dutch workshop realized that they are performing similar decision-making process as the "polder model," which is a Dutch version of consensus-based economic and social policymaking to look for a middle ground [54]. As D5 summarized: *"So there's an administrative question here. Which result do I want? Imagine that we're handling the polder model. We're starting at the average and then basing our discussion on that?"* D5 also urged his peers to consider what they can afford to sacrifice: *"We need to set priorities and make choices. And from the result you can see what we can afford to sacrifice"*. Similarly, at the end of the second English workshop, E6 synthesized what he discussed with E4 and E5 and concluded: *"I am okay with lowering my threshold a little bit considering the majority of people who might take a look at the Recent Changes, probably will not be experienced editors."*

Furthermore, we noticed that our **participants have been actively situating their collective decisions into their own community context.** That is, the Toolkit helped them go beyond merely choosing a machine learning model, probing them to think more about the purpose of the AI system as well as the larger community context within which this AI system is situated. For example, since the Dutch Wikipedia is relatively small and homogeneous, with almost all of its members living in the same time zone and interacting with each other both online and offline (see Table

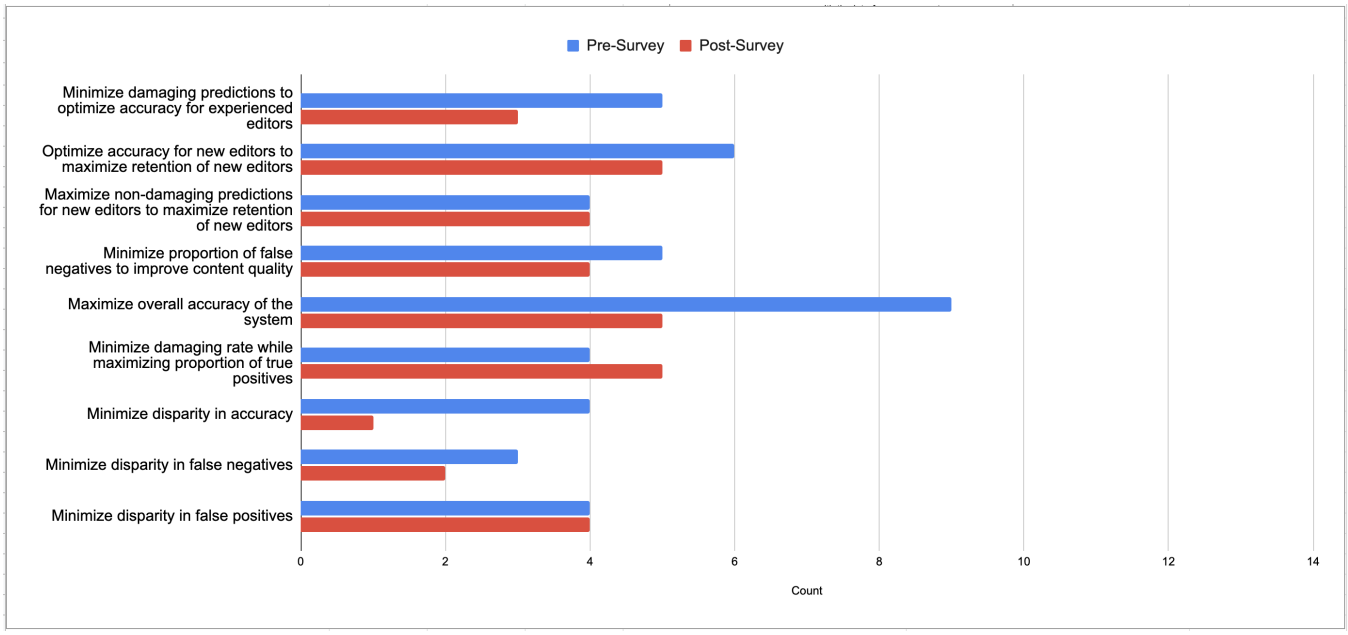


Figure 4: The Changes of Participants' Priorities in Maintaining the Content Quality of Wikipedia (N=14). The x-axis represents the count of votes participants give to each priority on the y-axis. Figure 4 shows that participants are less likely to prioritize overall accuracy in maintaining the content quality of Wikipedia after the workshops. In addition, it also suggests that participants narrow down their lists of priorities after deliberation workshops.

1), some participants questioned the need of deploying a machine-learning based system. In other words, they moved beyond from "choosing which model" to "whether do we need any model at all". D4 raised this question: *"In what capacity is this adding to what is already being done? For example, almost all of us are in the same time zone, if anyone posts anything at, say, 3am in the morning, we will know this one is problematic, right? Do we really need ORES?"*

In contrast, for the relatively larger and more diverse English Wikipedia (see Table 1), participants paid more attention on the issues of culture diversity. For instance, when asked to collectively deciding on high-level design principles, E1 added: *"They (bots) have to understand and capture those cultural variances. And that it's typically not easy"*. For example, E1 realized that if she wrote something about *"the culture of thummell or Tamil Nadu"*, ORES might not recognize many words or typical terms that belong to a particular community or culture. Therefor, E1 proceeded to propose that *"ORES system needs to understand and record various cultural differences and also address knowledge gaps"*.

Another example emerged when discussing the prioritized editor groups ORES should consider. Since Dutch Wikipedia has a relatively well-structured content moderation system consisted primary of human moderators (see Table 1), participants from the Dutch workshops tried to strive a balance between minimizing moderators' efforts and protecting newcomers. As D2 argued, although it is important to consider the workload of moderators, *"I have a feeling that we as a group are a bit different than the newcomers, so to say now that we should mark all of the newcomer edits immediately red to minimize moderators' effort is a bit too much? That's why I said I'd rather place a question mark here as a potential risk instead of looking at it as black or white... Don't be relentless, I don't want to scare off new editors"*.

In contrast, probably due to the large community size, a large number of new edits made everyday, and the relatively less well-structured content moderation system (see Table 1), English Wikipedia members prioritized *"reducing moderators' workload and assigning different kinds of maintenance tasks"* (E6) when the group were collectively reflecting on their design priorities. E4 seconded E6's comment on the importance of reducing workload of moderators and asserted that ORES design should avoid *"flooding them with too many things as their plates are probably already full"*.

Although at the end, only two of the four workshops came to a consensus on their chosen machine learning models, we observed that our Toolkit and workshops had nevertheless facilitated collective discussion and decision-making of system design within our participants' own community contexts.

5 DISCUSSION: TOWARD COMMUNITY-CENTERED, DELIBERATION-DRIVEN AI DESIGN

In this work, we presented the Model Card Authoring Toolkit to support community members navigate a spectrum of machine learning models and pick the ones that best align with their collective values. We hope that our study complements previous work in community-centered AI design by offering one possible solution to collective decision-making via *trade-off deliberation*. Below we discuss challenges, design opportunities, our limitations and future work.

5.1 Challenges and Design Opportunities

Our study reveals a few challenges and design opportunities in this emerging research area.

First, there are *time and resources challenges*. As Pierre et al. [50] reflected via their own research with local communities, data-centered participatory design projects can sometimes put a burden to community members, especially to marginalized social groups as they are often short on time and resources. This was reflected in our project. Indeed, we experienced difficulties recruiting participants for our project. Sometimes interested community members signed up for the workshop but eventually didn't make it due to time constraints. For example, they thought the workshop was too long and they didn't have enough time. Or, the proposed workshop time didn't fit into their schedule as they had to work or take care of family members. This suggests opportunities to conduct alternative types of deliberation. For example, instead of online synchronous workshops conducted in this study, we can also host asynchronous deliberation sessions through Wikipedia's internal forum such as Village Pump. As the majority of the exchanges between Wikipedia editors already take place on the Wikipedia site asynchronously, asynchronous deliberation might make it easier for editors from different time zones to engage in the conversation at the time that works best for them.

Second, there are *tech literacy challenges*. In our study, although we tried different ways of simplifying the toolkit via an iterative design process, including condensing the technical details of the models using the Model Cards approach [47], participants still spent a fair amount time trying to understand the terminologies used in the study. In particular, they had difficulties in understanding key performance metrics of machine learning models, such as false negatives and false positives. In one of our workshops, D5 had to explain the concept "false positives" to others as *"it's like you are tested positive for COVID but actually you don't have it"*. This suggests opportunities of further lowering enter barriers for community-centered AI design. For example, can we provide light-weight interactions to quickly contextualize the terminologies for non-expert community members so they can better understand performance metrics?

Finally, there are *infrastructural challenges*. This was especially present in our first workshop with members from the English Wikipedia as two of the participants were from Global South countries and their network connection was very unstable. In fact, they had to access Zoom via mobile phone networks due to the lack of cable in their communities. During our workshop, E2 even had to take a break to purchase new prepaid phone card to rejoin the meeting. As E2 commented, *"I hope future studies can consider the challenges we are facing. Not everyone has good networks or fancy computers."* This is a good reminder of the importance of underlying infrastructures supporting any forms of computer-mediated communication. One way of supporting online communities without robust network infrastructures is to conduct asynchronous forms of deliberation, as mentioned above. In addition, it also reminds us the importance of considering the diverse needs and constraints real-world community members can face in their everyday lives and how to better accommodate those needs and constraints in future work on community-centered AI design.

5.2 Limitations and Future Work

As a "proof-of-concept" case study, there are a number of limitations that are important to mention.

First, the online community we are working with in this project – Wikipedia – is a very special type of community. As an grass-roots online encyclopedia based on wiki technology, Wikipedia is often cited as a prominent example of the new community-based "flat" model for producing content and knowledge on the Internet [9]. Although studies have investigated the power dynamics within Wikipedia's governance model and organizational structures (e.g., [5, 62]), it nevertheless remains very different from real-world offline communities. For example, over the years, Wikipedia has developed a variety of deliberation mechanisms for resolving conflicts in content-related disputes [35], which offers a fruitful starting point for our workshops. Thus, how effective our approach will be in helping real-world, offline community members design their own decision-making algorithms remains as important question for future research.

Second, it is worth noting that community-based collective decision-making can be achieved via a number of ways. Deliberation is just one of them. Past scholarship has provided critiques of both the concept and the practice of deliberation (e.g., [25]). In this work, we used deliberation as one potential solution to support trade-offs discussions. Indeed, although our workshops have facilitated collective decision-making, only two out of four groups have reached a consensus. Future work is needed to explore alternative solutions. For instance, the theories of agnostic pluralism [48] illustrate the importance of contentious expression as a supplement and alternative form to deliberation, especially in communities with significant power dynamics. Future work should also explore other options in this domain.

Third, the research agenda of community-centered AI design demands long-term, deep engagement with real-world communities, to which our work is just a starting point. Future work is needed to support sustainable participation from a wide range of community members and to actually incorporate their decisions and feedback into the AI development pipeline. This requires new design approaches and mindsets [34], to which we hope to contribute in our future work.

6 CONCLUSION

In this paper, we present the Model Card Authoring Toolkit, a toolkit that supports community members to navigate a spectrum of machine learning models via deliberation and try to pick the ones that best align with their collective values. We documented an early use of our approach through a series of workshops in two different Wikipedia communities – English Wikipedia and Dutch Wikipedia. Our results suggested that the use of the toolkit is promising in helping community members understand, negotiate and collectively decide on the threshold for a machine learning based quality prediction system used in their communities' content moderation applications.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) under Award No. IIS-2001851, CNS-1952085, IIS-2000782, and the

NSF Program on Fairness in AI in collaboration with Amazon under Award No. IIS-1939606. Special thanks to Steven Wu, Xu Wang, Aaron Halfaker, Aditi Chattopadhyay, our colleagues at the HCII at Carnegie Mellon University, our anonymous reviewers, and all our participants.

REFERENCES

[1] Julia Abelson, Pierre-Gerlier Forest, John Eyles, Patricia Smith, Elisabeth Martin, and Francois-Pierre Gauvin. 2003. Deliberations about deliberative methods: issues in the design and evaluation of public participation processes. *Social Science & Medicine* 57, 2 (2003), 239–251.

[2] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[3] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 871–875.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] Ofer Arazy, Oded Nov, and Felipe Ortega. 2014. The [Wikipedia] world is not flat: On the organizational structure of online production communities. In *Proceedings of the Twenty Second European Conference on Information Systems*. 1–15.

[6] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the wikipedia language gap. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*. 1075–1084.

[7] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671–732.

[8] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, A Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[9] Yochai Benkler. 2008. *The wealth of networks*. Yale University Press.

[10] Dimitris Bertsimas, Arthur Delarue, William Eger, John Hanlon, and Sebastien Martin. 2020. Bus routing optimization helps Boston public schools design better policies. *INFORMS Journal on Applied Analytics* 50, 1 (2020), 37–49.

[11] Patrick Biernacki and Dan Waldorf. 1981. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research* 10, 2 (1981), 141–163.

[12] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[14] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 77–91.

[16] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 1–12.

[17] Robert J Cavalier. 2011. *Approaching deliberative democracy: Theory and practice*. Carnegie Mellon University Press.

[18] Kathy Charmaz. 2014. *Constructing grounded theory*. Sage.

[19] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 3539–3550.

[20] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[21] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 134–148.

[22] Sarah Dean, Sarah Rich, and Benjamin Recht. 2020. Recommendations and user agency: The reachability of collaboratively-filtered information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 436–445.

[23] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[24] James D. Fearon. 1998. Deliberation as discussion. In *Deliberative democracy*, J. Elster (Ed.). Cambridge University Press, Cambridge, MA, 44–68.

[25] Nancy Fraser. 1992. Rethinking the public sphere: A contribution to the critique of acutally existing democracy. In *Habermas and the public sphere*, C. Calhoun (Ed.). The MIT Press, Cambridge, MA, 109–142.

[26] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. 498–510.

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[28] Ben Green. 2021. Data science as political action: Grounding data science in a politics of justice. *Journal of Social Computing* 2, 3 (2021), 249–265.

[29] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[30] Aaron Halfaker and John Riedl. 2012. Bots and cyborgs: Wikipedia's immune system. *Computer* 45, 03 (2012), 79–82.

[31] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[32] Brent Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems*. 291–300.

[33] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[34] Yen-Chia Hsu, Ting-Hao 'Kenneth' Huang, Himanshu Verma, Andrea Mauri, Illah Nourbakhsh, and Alessandro Bozzon. 2022. Empowering local communities using artificial intelligence. *Patterns* 3, 3 (2022), 100449.

[35] Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. 2018. Deliberation and resolution on Wikipedia: A Case Study of Requests for Comments. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.

[36] Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel Procaccia, and Christos-Alexandros Psomas. 2019. Statistical foundations of virtual democracy. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR 97:3173–3182.

[37] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: Lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 45–55.

[38] Jon Kleinberg. 2018. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 40.

[39] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. 22–27.

[40] PM Krafft, Meg Young, Michael Katell, Jennifer E Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, et al. 2021. An action-oriented AI policy toolkit for technology audits by community advocates and activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 772–781.

[41] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory approaches to machine learning. In *International Conference on Machine Learning Workshop*.

[42] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: Protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 177–188.

[43] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[44] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[45] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[46] MediaWiki. 2021. ORES – MediaWiki. https://www.mediawiki.org/w/index.php?title=ORES&oldid=4901890

[47] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019.

Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.
[48] Chantal Mouffe. 1999. Deliberative democracy or agonistic pluralism? *Social research* (1999), 745–758.
[49] Otter.ai. 2020. Otter.ai, Inc. https://otter.ai
[50] Jennifer Pierre, Roderic Crooks, Morgan Currie, Britt Paris, and Irene Pasquetto. 2021. Getting Ourselves Together: Data-centered participatory design research & epistemic burden. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.
[51] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.
[52] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. 2021. Towards fairness in practice: A practitioner-oriented rubric for evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[53] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[54] Yda Schreuder. 2001. The Polder model in Dutch economic and environmental planning. *Bulletin of Science, Technology & Society* 21, 4 (2001), 237–245.
[55] Hetan Shah. 2018. Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 20170362.
[56] Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 850–861.
[57] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.
[58] Hong Shen, Ángel Alexander Cabrera, Adam Perer, and Jason Hong. 2020. "Public(s)-in-the-Loop": Facilitating Deliberation of Algorithmic Decisions in Contentious Public Policy Domains. In *Fair & Responsible AI Workshop CHI2020*.
[59] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a design fix for machine learning. In *Proceedings of the International Conference on Machine Learning*. 1–7.
[60] 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*. 1–14.
[61] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).
[62] Laura Stein. 2013. Policy and participation on social media: The cases of YouTube, Facebook, and Wikipedia. *Communication, Culture & Critique* 6, 3 (2013), 353–371.
[63] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 10–19.
[64] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.
[65] Miaomiao Wen, Keith Maki, Steven Dow, James D Herbsleb, and Carolyn Rose. 2017. Supporting virtual team formation through community-wide deliberation. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.
[66] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The What-If tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 56–65.
[67] Pak-Hang Wong. 2020. Democratizing algorithmic fairness. *Philosophy & Technology* 33, 2 (2020), 225–244.
[68] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[69] Yiwei Yang, Eser Kandogan, Yunyao Li, Prithviraj Sen, and Walter S Lasecki. 2019. A study on interaction in human-in-the-loop machine learning for text analytics. In *Joint Proceedings of the ACM IUI 2019 Workshops*.
[70] Zining Ye, Xinran Yuan, Shaurya Gaur, Aaron Halfaker, Jodi Forlizzi, and Haiyi Zhu. 2021. Wikipedia ORES Explorer: Visualizing Trade-offs For Designing Applications With Machine Learning API. In *Proceedings of the 2021 ACM on Designing Interactive Systems Conference*. 1554–1565.
[71] 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 on Designing Interactive Systems Conference*. 1245–1257.
[72] 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), 1–23.