Analyzing Biases at Multiple Stages of an Algorithmic Decision Support System

Xingyu Liu
Carnegie Mellon University
Pittsburgh, USA

Zhiwei Steven Wu
University of Minnesota
Minneapolis, USA

Haiyi Zhu
Carnegie Mellon University
Pittsburgh, USA

ABSTRACT

With the rapid growth of Machine Learning-based decision support systems, it has become more and more important for both developers and users to be aware of, understand and mitigate bias in these systems. In this paper, we discuss our on-going study to examine the biases in ORES, a community-developed and -maintained vandalism fighting system. We provide a flow model that captures its entire process from data collection to deployment. We break down the flow model into four phases and study how different forms of biases can arise and propagate throughout the ORES pipeline. Specifically, we (1) identify potential sources of biases across four phases of the pipeline, (2) show evidence of biases through quantitative and qualitative analysis, and (3) propose interventions to mitigate these biases and test their feasibility within the real-world context. Finally, we provide a checklist of questions for both developers and users to discover and locate bias and fairness issues at each stage of the system, and also directions for future research in fairness in Machine Learning and HCI.

Author Keywords

Bias; Fairness; Machine Learning; Decision Support System; Wikipedia

INTRODUCTION

Increasingly, AI technologies have been used to allocate, optimize and evaluate work of human in a wide range of work domains, ranging from traditional workers such as Starbucks barista, mail deliverymen and warehouse workers, to workers on online platforms like Uber, Lyft, TaskRabbit and Mechanical Turk. Wikipedia, which is the largest peer production project in human history, has been increasingly relying on intelligent tools, ranging from machine learning based service, fully automated agents, to semi-automated tools, to manage workflow and evaluate contributors' work [17, 13, 31, 6]. In particular, Wikipedia has developed a system of "surveillance" [9], carried out by human and AI-based tools, to fight massive vandalism on Wikipedia.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

However, concerns have been raised that the surveillance system on Wikipedia might be systematically biased: treating editors differently based on the groups they belong to, such as whether they are anonymous (without a registered account) or not, or whether they are newcomers or not [15, 8, 9, 24]. Scholars like de Laat argue that this type of differential treatment is unethical, immoral, and inconsistent with the community's founding principles of transparency and equality [8, 9]. Evidence has also revealed that the bias has hindered the long-term growth of the community [15]. Good-intentioned newcomers, who are still learning how to contribute, tend to leave the community when they are rudely greeted by the system who is supposed to fight vandals [16].

In this paper, we aim to explore and study how biases are introduced and propagated in Wikipedia's vandalism fighting system. We are currently conducting an in-depth case study of the ecosystem of Wikipedia's Objective Revision Evaluation Service (ORES). ORES is an open-source machine learning service designed to generate real-time predictions on edit quality and article quality, operating on 38 language versions in Wikipedia [14]. It is internally a collection of machine learning classifiers trained on manually labelled ground-truth data sets. The classifiers are used to predict different characteristics of newly added Wikipedia edits, including article quality, damaging or not, good-faith or not, etc. The prediction results are further incorporated into different decision support tools including the Recent Changes page[1], the Watchlist page[2], and Huggle[3], to help human patrollers to make decisions.

We are taking a holistic approach to examine the whole pipeline of the vandalism fighting system in the Wikipedia community. We will study each stage of the system, ranging from data collection, model training and evaluation, tool design, and human interaction (see Figure 1). We will examine the complex causes and issues of biases in each stage, and propose feasible recommendations that might mitigate biases.

Our work contribute to the rich ongoing conversation concerning the use of algorithms in decisions in our society. In contrast to a lot of research in this field focusing on the machine learning model part, we attempt to study the entire pipeline. Vandalism fighting in Wikipedia provides a real-world setting with lower stakes and detailed data. Our findings and approaches might be useful to understand biases and identify solutions for high-stake settings such as criminal justice.

[1]Recent Changes page: https://en.wikipedia.org/wiki/Special:RecentChanges
[2]Watchlist page: https://www.mediawiki.org/wiki/Special:Watchlist
[3]Huggle: https://en.wikipedia.org/wiki/Wikipedia:Huggle

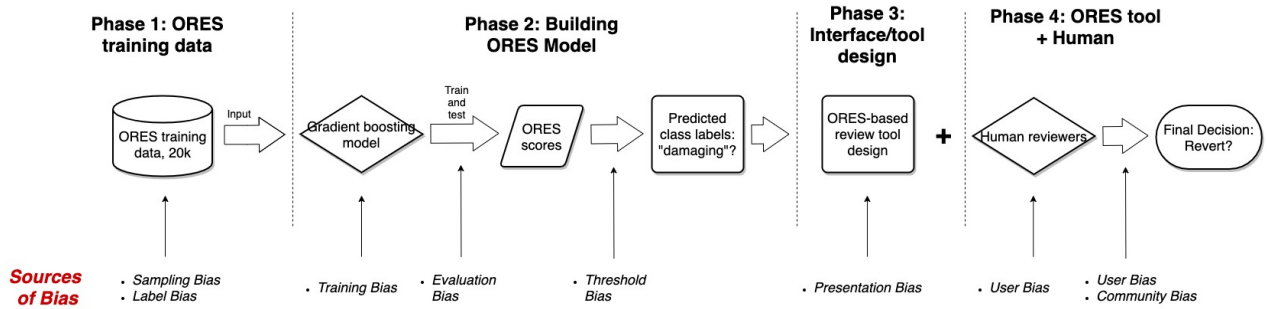

Figure 1. Our proposed flow model of how biases are introduced and propagate through the development and deployment of ORES.

SCOPE OF OUR RESEARCH

In the following section, we will define and clarify some important concepts and assumptions that we will use throughout this research.

Definition of Bias

We focus on the group-based bias in the ORES system. Formally, we define bias as the **differential treatment** between protected groups and unprotected groups. Following prior work [24], we define the protected groups are anonymous editors (who act on Wikipedia without logging in), and newcomers (who act on Wikipedia with a newly-created account). This definition is quantified and defined more formally in our analysis of the first two phases, and considered more broadly in the latter two phases. In the first two phases, bias is measured as the disparity of False Positive Rate (FPR) and False Negative Rate (FNR) on algorithm's prediction of "damaging" between newcomers, anonymous editors and experienced editors. In the ORES context, FPR represents the proportion of non-damaging edits that are falsely predicted as damaging in the group, while FNR represents the proportion of damaging edits that are falsely predicted as non-damaging in the group.

Type of Machine Learning Task

ORES is a set of machine learning models trained using a supervised classification approach, with explainable features and human-labeled ground-truth. ORES predictions are incorporated in many decision support tools. In most cases, humans make the final decision. Some of our takeaways and insights might not apply to other types of machine learning-based systems, such as reinforcement learning, unsupervised learning, convolutional neural networks, etc.

Multi-Cultural Contexts

Wikipedia contains 285 different languages versions [29], and each language has its own Wikipedia community with distinct community cultures and values. Thus, ORES is a service that support 38 languages [28] where each wiki language uses its own scale, with different training data, machine learning model, threshold settings, interface design etc. Reviewers in different wiki languages might also have different cultural values regarding the use of machine-supported decision making tool and biases and fairness issues.

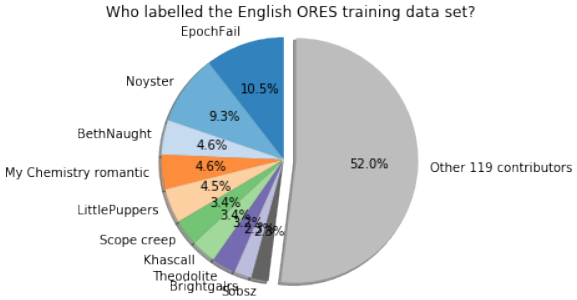

Figure 2. A pie chart showing the distribution of amount of contribution to the English ORES training data labelling. The top 10 contributors provided nearly half of all the labels.

In this work-in-progress paper, we focus on studying ORES in one specific (also the largest) Wikipedia community, the English Wikipedia community. Our long-term plan is to conduct a multiple-case study to audit ORES deployed in different language communities. We are interested in examining how each of the stages we proposed is different in different Wikipedia communities, how different states of each stage operate and lead to different results, and how different values and cultures cause different bias and fairness problems in this algorithmic decision support system.

CASE STUDY: BIASES IN ORES

In this section, we will present our preliminary findings on biases in the ORES system in the English Wikipedia community. We will walk through four phases (Figure 1): ORES training data, ORES model, interface and tool design, and human's interaction with ORES. For each phase we show evidence of biases, identify possible sources of biases, and explore current or proposed methods to mitigate bias.

Phase 1: Collecting ORES Training Data

There is by now a vast literature on how machine learning models can inherit biases from their training data (see e.g. [5] for a survey). In the context of ORES, we focus on two potential sources of bias that can come from labeling and sampling processes.

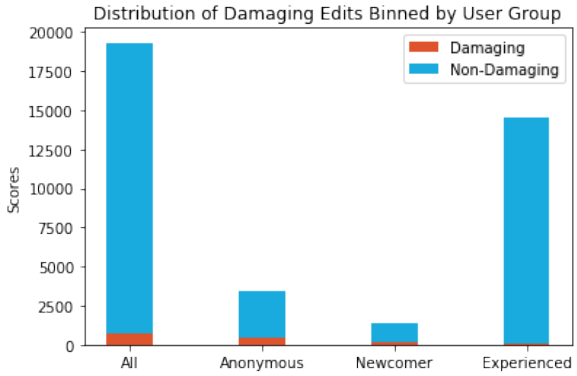

Figure 3. A stacked bar chart displaying the distribution of anonymous, new-coming and experienced editors and proportions of damaging edits within each user group in the training data set of English ORES.

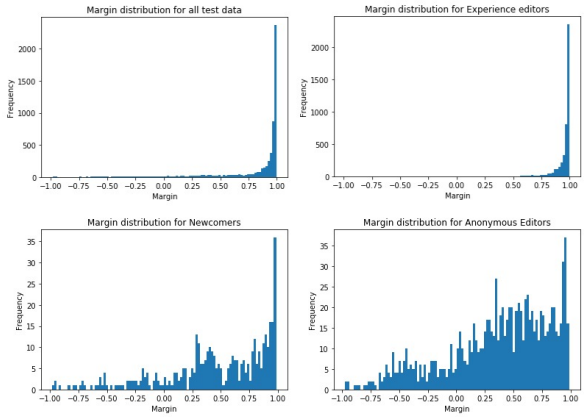

Figure 4. Histograms of margin distribution of all editors, experienced editors, newcomers and anonymous editors.

Label Bias

Labels in the ORES training data are manually collected through Wiki Labels[4], a crowdsourcing platform developed by Wikimedia Foundation. In Wiki Labels, each data point is only labelled by one volunteer, with no review process. In addition, in four years of deployment, only 129 volunteers were involved in this labelling process, with 48% of all the data labelled by the top 10 contributors (Figure 2). Possible biases could derive from this small group of Wikipedia enthusiasts. For example, they may hold a higher standard of quality and may be overly strict towards newcomers' edits.

To mitigate label bias, developers could encourage a diverse participation of the data labelling process, and incorporate a review/approval step in the crowdsourcing pipeline [27, 22].

Sampling Bias

Sampling bias arises when data is not collected randomly and uniformly. There are two kinds of sampling bias in ORES. First, ORES's training data is not uniform across different user groups. Among 20,000 data points, 72.45% of the edits are made by experienced users; 17.37% of the edits are made by anonymous editors; and only 10.18% of the edits are made by new-coming editors. Secondly, the base rates of training examples with "damaging" labels are very different across the three groups. In the training data set, 13.7% of anonymous users' edits are labelled as damaging; 14.5% of newcomers' edits are labelled damaging; but only 0.05% of experienced users' edits are labelled as damaging (Figure 3). Similar imbalanced distribution also exist for the "good-faith" label.

Mitigating sampling bias is well-studied in the machine learning community, as many researchers consider it as the main source of bias. One approach is to massage the data to remove bias. Examples include [4, 20, 30, 11]. Another way of bias mitigation is to re-weight the data points without changing the labels, for example [19, 21]. We experimented with a simple and intuitive method of balancing the proportion of damaging edits for each user group in the training data, and train a machine learning model with exactly the same hyper-parameters. We observed that the disparity of FPR and FNR between user groups are significantly reduced comparing to the original ORES model, while still maintaining a relatively high balanced accuracy (Table 1).

[4]Wiki Labels: https://labels.wmflabs.org/stats/enwiki/41

Phase 2: Building ORES Model

Biases could be either inherited and amplified from the data or newly generated in the model building process. In this phase, we identify three possible sources of bias in the ORES context: training bias, evaluation bias and threshold bias.

Training Bias

In general, the machine learning training algorithm optimizes the overall predictive accuracy based on the training data without explicitly accounting for fairness criteria, which can lead to different forms of discrimination, such as large disparity of FNR and FPR (see Table 1). There has been recent work on how to enforce different fairness notions in the training process (see e.g. [18, 1]).

Evaluation Bias

Evaluation bias happens during the model evaluation process. This include inappropriate selection of performance metrics, or imbalanced distribution of the test data set. ORES's disparity of treatment to anonymous editors and newcomers was hidden in the model evaluation step, because the test data set of ORES was highly imbalanced with respect to user group (most of the edits are from experienced editors).

Threshold Bias

In many methods for binary classification, the predictor provides a probability estimate and predicts the binary label by thresholding the probability estimate. The standard threshold are set default to 0.5. However, different threshold settings might produce different bias level in the system. In figure 4, we demonstrate how different threshold settings might lead to disparity in treatment (in this case, FPR and FNR) between anonymous, newcomer and experienced editor groups. In machine learning, margin measures the confidence/uncertainty level of a single prediction made by the model. Formally it is

	Models of ORES		
Metric	*Original*	*Balanced labels for each group*	*Removed protected features*
Balanced Accuracy	0.749	0.714	0.716
F1 Score	0.427	0.497	0.368
FPR Anonymous	0.165	0.156	0.118
FPR Newcomer	0.116	0.120	0.072
FPR Experienced	0.000	0.092	0.019
FNR Anonymous	0.405	0.423	0.504
FNR Newcomer	0.467	0.424	0.578
FNR Experienced	0.944	0.522	0.611

Table 1. Table comparing performance and bias level of (1) the original ORES model, (2) a new model trained with balanced proportion of damaging edits withing each user group, and (3) a model trained on data with protected features dropped. This shows the effect of an experiment mitigating sample bias and feature bias.

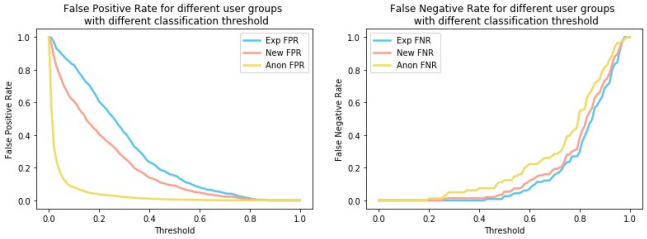

Figure 5. Two line charts showing the FPR and FNR of different user groups at different threshold settings of classification.

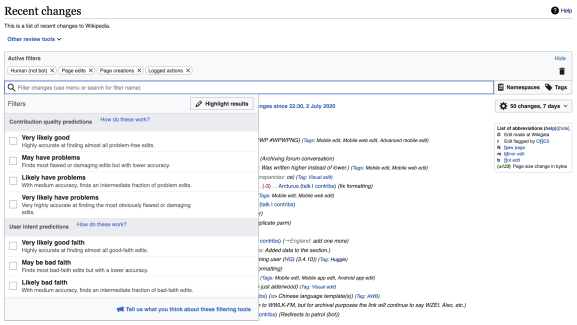

Figure 6. The interface of the Recent Changes page, which inform users the prediction results from ORES by color-coding and filtering.

computed as:

$$\text{margin}_i = y_i \times f(x_i), \tag{1}$$

where y_i is the true label (either -1 or 1) of the ith data, $f(x_i)$ the model's prediction score of the ith data (within the range of -1 to 1). Thus, a confident correct prediction should have a margin score around 1, a confident but wrong prediction should have a margin score around -1, and an uncertain prediction should have a margin score around 0.

Figure 4 shows that the margin distributions are vastly different for different user groups. Thus setting different thresholds for classification will have significant impact on the disparity of FPR and FNR. This effect is further illustrated in figure 5, which displays the FPR and FNR of each user group at different threshold levels. We can observe that disparities of FPR and FNR between different user groups (i.e. gaps between the lines at the same point on x-axis) are larger at some threshold levels.

Phase 3: Interface/tool Design

Algorithmic decision support systems like ORES are driven by the internal algorithm. However, machine learning models only provide signals. In the case of ORES, model outputs a prediction score at some confidence level, which is later classified by a threshold. A user interface or tool is required to connect the raw scores or classifications to a real-world reviewer's usage, and this interface and tool design can greatly affect how people understand, interact with, and make final decisions based on ORES. There are 38 existing tools[5] built using ORES scores, and among those the two most popular tools are the Recent Changes page[6] and Huggle[7].

Presentation Bias

Different ways of presentations might produce biases. The Recent Changes page employed two different ways of presenting prediction results from ORES: filtering and color-coding (Figure 6). Users can filter and/or color-code edits based on their level of "probability to be damaging". TeBlunthuis et al. discovered that edits flagged on the recent changes page are reverted more often, especially edits made by experienced editors, thus leading to a increase in fairness [24]. However, this research was treating the filtering and color-coding functions as a whole, and only approximate bias as the "frequency of being revert of each user group". In addition to filtering and color-coding predicted damaging edits, Huggle used a ranking system to present ORES's predictions.

We are planning to conduct a factorial experiment to examine how these three presentation methods, filtering, color-coding and ranking, might result in an increase/decrease of the fairness of this decision support tool.

[5]Tools built on ORES: www.mediawiki.org/wiki/ORES/Applications
[6]Recent Changes page: en.wikipedia.org/wiki/Special:RecentChanges
[7]Huggle: en.wikipedia.org/wiki/Wikipedia:Huggle

Phase 4: Human Interacting with Interface/Tool

Human are imperfect. When ORES and ORES-based interfaces and tools are used to support decisions that are ultimately made by humans, the algorithms' and systems' fairness or biases are not guaranteed to carry over to decision outcomes.

User Bias

A long line of social science research has shown that users have cognitive and social limitations when working with automated or intelligent systems. For example, some users may distrust decision-support systems and choose to completely ignore algorithmic suggestions. This is often called "algorithm aversion," a phenomenon where human users are reluctant to use algorithms, even when they are proven to be more accurate than human judgment [12, 10]. Conversely, some users may rely on automated system too much, and form an automation bias, accepting recommendations mindlessly and failing to question the algorithms even when algorithms make errors[7]. Furthermore, studies show that people rely on a limited number of heuristics to assess probabilities, which leads to severe and systematic errors [25]. A series of studies demonstrated that humans mindlessly apply social rules and expectations to automated or intelligent systems. A recent study [3] shows that when physicians are first introduced to a diagnostic AI assistant, they want to know the system's subjective "point of view," such as the extent to which it tends to be more liberal or conservative when grading cancer severity. Last, different people may exhibit different levels and types of cognitive limitations and biases.

However, there is limited understanding of how human reviewers use the ORES powered interface and tools, and to what extent the final decisions are biased. In future work, we want to answer the following questions: what behavioral patterns do reviewers exhibit when they interact with ORES-based interface and tools? How do the reviewers' behaviors and workflows impact the bias and fairness of the final decisions? To answer these questions, we will first observe and conduct think-aloud exercises and contextual interviews with reviewers; we will then conduct field experiments to further quantify the user biases.

Community Bias

English Wikipedia community does not perceive all members of the community as equally trustworthy[23]. For example, it implements higher trust user groups, such as "Bureaucrats" who can grant user group permissions, and "Administrators" who have additional controls over pages and other users' accounts. In contrast, anonymous editors who act without accounts or newcomers who act with new accounts, are often perceived as less trustworthy by many community members. The emphasis of "trustworthiness" of the person in the English Wikipedia community influences how reviewers make decisions, especially when they are dealing with ambiguous cases from newcomers and anonymous editors.

In the future study, we want to compare different language versions of Wikipedia and examine how the community norms and cultures might influence the bias of vandalism systems.

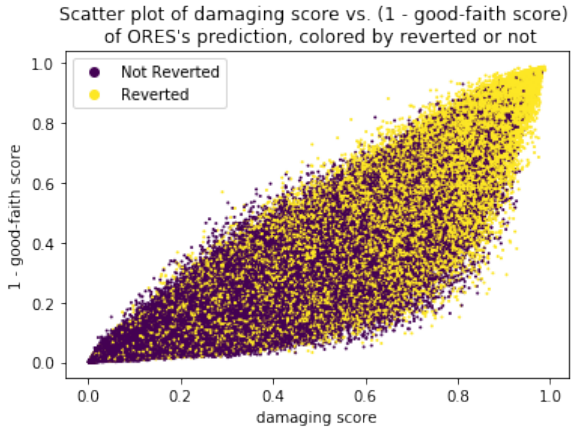

Figure 7. A scatter plot of 1 - good-faith score vs. damaging score predicted by ORES, colored by if the edit is reverted or not. There are no clear correlation between reverting and good-faith or damaging.

Problems Above All Phases

In addition to the biases that occur in the four phases described above, we also identified several bias and fairness problems that lay above all phases.

Mismatch between machine learning task and human task

The ORES system was trained to predict whether an edit is "damaging" or not, and "good-faith" or not. However, Wikipedia reviewers only take that as an indicative factor to consider, and their final task is to determine whether an edit should be reverted or not. Even if a reviewer has unbiased understanding of all information provided, reverting is still a complex behavior which might be influenced by reviewers' different levels of enthusiasm on a topic, communication error or incorrect knowledge etc. In figure 7, we show the relationship between the final result of reverted or not vs. the two predicted scores by ORES ("damaging" and "good-faith"). While most of the edits at the two extremes (high damaging score and low good faith score or vice versa) are reverted (or not reverted in the vice versa case), the data is particularly noisy in the middle part of the distribution.

Unclear definition of "damaging" and "should be reverted"

Many of machine learning classification problems have clear definitions of labels. However, in the context of Wikipedia, it is hard to provide a clear definition of what edits are damaging and should be reverted. Even if rules are regulated, people's different understandings and standards might lead to noisy results.

Arbitrary classification of user groups

While many of the fairness research have clearly defined protected features, in Wikipedia we have to arbitrarily bin users into groups (Newcomer vs. Experienced editors), based on a continuous feature of their time on Wikipedia. In this paper, we first identify anonymous editors (and registered editors) by their registration status when making the edit. We then quantitatively defined newcomers as users who have the *seconds_since_registration* attribute value in the lower 25 per-

centile (less than 3.63×10^6 seconds, approximately 42 days) of registered users, and other users as experienced editors.

Non-observable protected features

Some of the protected attributes are particularly hard to observe on Wikipedia, thus, it is difficult to audit bias issues with respect to these features. For example, imbalanced proportion of male and female editors has long been a problem for the Wikipedia community. In particular, the 2010 UNU-MERIT survey [26] found evidence of a significant gender skew: fewer than 13% of Wikipedia contributors are women. However, research found that the amount of male and female new-coming editors are approximately the same [2]. One hypothesis that explains the extremely low retention rate of female editors is that female editors' edits are more likely to be reverted (and falsely reverted) comparing to male editors, which significantly discourage involvement. However, this problem remains unsolved because of the non-observable nature of the gender attribute.

DISCUSSION

Checklist of Questions

Based on our exploratory audit of bias and fairness issues in ORES, we propose a generalizable checklist of questions that could be used by developers and users of similar algorithmic systems, especially people who are less familiar with machine learning and bias and fairness issues, to discover and locate sources of biases. Each question correspond to one potential bias in the flow model we discussed:

1. How are the labels of data defined? Are there clear definitions of labels? Do contributors have consistent understandings of labels?
2. How was the training data collected? Who labelled the data?
3. Is there a large disparity on the ratios of positive/negative examples across different groups?
4. Does the trained model violate certain statistical fairness notions?
5. Is the test data balanced w.r.t. the protected features? What metrics are used in testing?
6. What threshold is set for model's classification from output scores?
7. How are the classification results presented?
8. Who are the general users of this tool? What are their bias and values?

Future Work

In addition to our early stage findings described in this paper, we will conduct more formal and detailed research in four aspects: (1) Since different Wikipedia communities have separate development, deployment and usage of ORES, we will do a multiple-case study to analyze the variability of our proposed bias flow model. For example, how different states of each stage operate and lead to different results. (2) We will conduct think-aloud and semi-structured interviews to understand human's real-world interaction with ORES, their believes and values, and qualitatively analyze biases and fairness issues in phase 3 and 4. (3) To explore how different presentation and interaction methods influence bias level in algorithmic decision support systems, we will conduct a factorial experiment with conditions of different interface designs, including ranking, filtering, highlighting etc. (4) To evaluate the generalizability of our bias flow model, we will conduct a agent-based modeling study to help us understand and quantify factors that lead to biases in algorithmic decision support systems

We also identified several possible directions for future bias and fairness research in Machine Learning and HCI. First, we noticed that most bias and fairness research focus on the machine learning model part with benchmark datasets like COMPAS, DiF, and UCI Adult Dataset, and few examine the real-world workflow, neglecting phase 3 and 4. Secondly, most of the current bias mitigation strategies are addressing bias either from the developer's side who have access to the internal data and algorithm, or require professional knowledge of machine learning and algorithmic bias to build a post-processing model. Few research have discussed methods to allow everyday users of the system, who are not machine learning experts to mitigate biases on their own, especially for black box systems.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[2] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. 2011. Gender Differences in Wikipedia Editing. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. Association for Computing Machinery, New York, NY, USA, 11–14. DOI: http://dx.doi.org/10.1145/2038558.2038561

[3] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops (ICDMW '09)*. IEEE Computer Society, USA, 13–18. DOI:http://dx.doi.org/10.1109/ICDMW.2009.83

[5] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. *CoRR* abs/1810.08810 (2018). http://arxiv.org/abs/1810.08810

[6] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*. 32–41.

[7] Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*. 6313.

[8] Paul B de Laat. 2015. The use of software tools and autonomous bots against vandalism: eroding Wikipedia's moral order? *Ethics and Information Technology* 17, 3 (2015), 175–188.

[9] Paul B de Laat. 2016. Profiling vandalism in Wikipedia: A Schauerian approach to justification. *Ethics and Information Technology* 18, 2 (2016), 131–148.

[10] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[11] Michael Feldman. 2015. Computational Fairness: Preventing Machine-Learned Discrimination.

[12] Robert Fildes and Paul Goodwin. 2007. Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces* 37, 6 (2007), 570–576.

[13] R Stuart Geiger and Aaron Halfaker. 2013. When the levee breaks: without bots, what happens to Wikipedia's quality control processes?. In *Proceedings of the 9th International Symposium on Open Collaboration*. 1–6.

[14] Aaron Halfaker and R.Stuart Geiger. 2019. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. (09 2019).

[15] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist* 57, 5 (2013), 664–688.

[16] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*. 163–172.

[17] Aaron Halfaker and John Riedl. 2012. Bots and cyborgs: Wikipedia's immune system. *Computer* 45, 3 (2012), 79–82.

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 3315–3323. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning

[19] Heinrich Jiang and Ofir Nachum. 2020. Identifying and Correcting Label Bias in Machine Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Silvia Chiappa and Roberto Calandra (Eds.), Vol. 108. PMLR, Online, 702–712. http://proceedings.mlr.press/v108/jiang20a.html

[20] F. Kamiran and T. Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. 1–6.

[21] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-Aware Classification. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 853–862. DOI: http://dx.doi.org/10.1145/3178876.3186133

[22] Brian Lee, Sulin Ba, Xinxin Li, and Jan Stallaert. 2018. Salience Bias in Crowdsourcing Contests. *Information Systems Research* 29 (03 2018). DOI: http://dx.doi.org/10.1287/isre.2018.0775

[23] 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.

[24] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. 2020. The effects of algorithmic flagging on fairness: quasi-experimental evidence from Wikipedia. *arXiv preprint arXiv:2006.03121* (2020).

[25] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.

[26] UNU-MERIT. 2010. Wikipedia Survey – Overview of Results. (2010). https://www.merit.unu.edu/wp-content/uploads/2019/03/Wikipedia_Overview_15March2010-FINAL.pdf.

[27] Fabian L Wauthier and Michael I. Jordan. 2011. Bayesian Bias Mitigation for Crowdsourcing. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1800–1808. http://papers.nips.cc/paper/4311-bayesian-bias-mitigation-for-crowdsourcing.pdf

[28] Wikipedia. 2020a. List of ORES models in wiki languages. (2020). https://github.com/wikimedia/editquality/tree/master/models.

[29] Wikipedia. 2020b. List of Wikipedias. (2020). https://en.wikipedia.org/wiki/List_of_Wikipedias.

[30] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Sanjoy Dasgupta and David McAllester (Eds.), Vol. 28. PMLR, Atlanta, Georgia, USA, 325–333. http://proceedings.mlr.press/v28/zemel13.html

[31] 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.