Value-Sensitive Algorithm Design: Method, Case Study, and Lessons

HAIYI ZHU, University of Minnesota, Twin Cities
BOWEN YU, University of Minnesota, Twin Cities
AARON HALFAKER, Wikimedia Foundation
LOREN TERVEEN, University of Minnesota, Twin Cities

Most commonly used approaches to developing automated or artificially intelligent algorithmic systems are Big Data-driven and machine learning-based. However, these approaches can fail, for two notable reasons: (1) they may lack critical engagement with users and other stakeholders; (2) they rely largely on historical human judgments, which do not capture and incorporate human insights into how the world can be improved in the future. We propose and describe a novel method for the design of such algorithms, which we call Value Sensitive Algorithm Design. Value Sensitive Algorithm Design incorporates stakeholders' tacit knowledge and explicit feedback in the early stages of algorithm creation. This increases the chance to avoid biases in design choices or to compromise key stakeholder values. Generally, we believe that algorithms should be designed to balance multiple stakeholders' needs, motivations, and interests, and to help achieve important collective goals. We also describe a specific project "Designing Intelligent Socialization Algorithms for WikiProjects in Wikipedia" to illustrate our method. We intend this paper to contribute to the rich ongoing conversation concerning the use of algorithms in supporting critical decision-making in society.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**;

Additional Key Words and Phrases: Online Communities; Online Recruitment; System Buildings; Algorithmic Intervention; Value-Sensitive Algorithm Design; Peer Production; Wikipedia; WikiProjects

ACM Reference format:
Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 194 (November 2018), 23 pages.
https://doi.org/10.1145/3274463

1 INTRODUCTION

Automated and artificially intelligent algorithmic systems assist humans in making important decisions across a wide variety of domains. Examples include: helping judges decide whether defendants should be detained or released while awaiting trial [11, 42], assisting child protection agencies in screening referral calls [10], and helping employers filter job resumes [53]. Intelligent algorithms are also used to "govern" digital worlds. For example, in Wikipedia, various sophisticated tools are used to automatically assess the quality of edits and to take appropriate actions such as reverts [24].

Most of these decision-making or decision-supporting algorithms are developed using machine learning-based approaches, with a training process that *automatically* mines patterns from historical data. However, there is an emerging body of literature identifying ways that this approach might fail. First, automation may worsen engagement with key users and stakeholders. For instance, a series of studies have shown that even when algorithmic predictions are proved to be more accurate than human predictions, domain experts and laypeople remain resistant to using the algorithms

2573-0142/2018/11-ART194 $15.00
https://doi.org/10.1145/3274463

[15, 16, 45]. Second, an approach that largely relies on automated processing of historical data might repeat and amplify historical stereotypes, discriminations, and prejudices. For instance, African-American defendants were substantially more likely than Caucasian defendants to be incorrectly classified as high-risk offenders by recidivism algorithms [2], and Google Ads displayed more high-paying jobs for male users than for female users [14]. These automated algorithmic tools can cause unintended negative societal and community impacts. For instance, good-faith newcomers left the Wikipedia community in droves after their edits were reverted by algorithmic tools [29].

One possible way to address these issues is to increase human influence on automated algorithmic systems. Researchers have proposed frameworks like "human-in-the-loop" [64] and "society in the loop" [34, 57], suggesting that the judgment of individuals or the society as a whole should be embedded into automated algorithmic systems. In this paper, we propose a novel method called **Value-Sensitive Algorithm Design**, which engages relevant stakeholders in the *early* stages of algorithm creation and incorporates stakeholders' tacit values, knowledge, and insights into the abstract and analytical process of creating an algorithm. Value-Sensitive Algorithm Design aims to use human insights to guide the creation of automated algorithms, with the goal of increasing stakeholder acceptance and engagement, reducing potential biases in design choices, and preventing compromise of important stakeholders' values.

To illustrate the method, we describe in detail a case study we conducted in English Wikipedia to develop an algorithmic recruitment system for WikiProjects. WikiProjects are self-organized groups of Wikipedia contributors working together to improve coverage of specific topics. They serve as important hubs for editors to seek help and find collaborators, guide and organize editors' work, and offer protection against unwarranted reverts and edit wars [20, 78]. Despite the benefits WikiProjects provide, participation has been declining. While the number of WikiProjects reached a peak of about 2,157, as of July 2017 there were only 506 active WikiProjects, and only 12% of them had more than 10 active members[1]. If WikiProjects can serve as a socialization hub, this benefits Wikipedia as a whole, since socializing and retaining editors is a well-recognized problem [27]. Thus, our system aims to help WikiProjects identify and recruit suitable new members.

We followed our 5-step approach Value-Sensitive Algorithm Design method to create a value-sensitive recruitment algorithm for WikiProjects:

(1) We review literature and conduct empirical studies to understand relevant community stakeholders' motivations, values, and goals, and identify potential trade-offs.
(2) Based on the results of the first step, we identify algorithmic approaches and create prototype implementations.
(3) We engage and work closely with the community to identify appropriate ways to deploy our algorithms, recruit participants, gather feedback, etc.
(4) We deploy our algorithms, gather stakeholder feedback, and refine and iterate as necessary.
(5) We evaluate stakeholders' acceptance, algorithm accuracy, and impacts of the algorithms on the community.

Findings from our nine-month design, deployment, iteration, and evaluation process include:

- Our algorithmic tool was well received by the community;
- 16 organizers representing 18 WikiProjects actively used the tool. During the six-month period, these organizers received 385 newcomer recommendations and sent out 100 invitations.
- Inexperienced editors and experienced editors were equally invited by the project organizers through the recruitment algorithms.

[1]We consider a WikiProject active if there were more than 10 edits made on any related project pages or talk pages of that WikiProject in the previous month, and an editor active if the editor made more than 5 edits in the previous month.

- Experienced newcomers who received invitations from project organizers had a significant increase in their within-project activities over the baseline group.

After presenting the case study, we reflect on the lessons and challenges for the Value-Sensitive Algorithm Design method that the case study helped reveal. The fundamental goal of our method is to contribute to the ongoing conversation concerning the use of algorithms in supporting critical decision-making in our society.

2 RELATED WORK

As algorithms have become increasingly embedded throughout society, evidence has emerged suggesting that many of them have normatively problematic outcomes [62]. For example, racial minorities might be less likely to find housing via algorithmic matching systems [18]; algorithmically-controlled personalized job matching algorithms might restrict the information available for use by the economically disadvantaged [71]; online markets might unfairly make goods more expensive for particular demographics or particular geographic locations [72]. Studies also suggest evidence of racial discrimination in recidivism prediction algorithms [2] and gender bias in Google ads [14].

Researchers have called for the development of systematic methods to detect and fix bias and in existing algorithms [62]. Fairness-aware (or discrimination-aware) machine learning research attempts to translate non-discrimination notions mathematically into formal constraints and develop models that take such constraints into account. Zliobaite [80] summarizes three main approaches to applying constraints in algorithmic systems: training data pre-processing, model post-processing, and model regularization. Data pre-processing modifies historical data such that it no longer contains unexplained differences across protected and unprotected groups [19, 25, 36, 38, 47]. Model post-processing produces a standard model and then modifies this model to obey non-discrimination constraints [7, 26, 37]. Model regularization forces non-discrimination constraints during the model learning process [6, 37, 39].

Despite the mathematical rigor of these approaches, the results of an interview study with 27 public sector machine learning practitioners across 5 OECD countries [73] suggest a disconnect between the current discrimination-aware machine learning research and organizational and institutional realities, constraints and needs; this disconnect is likely to undermine practical initiatives. Researchers have proposed frameworks like "society in the loop" [34, 57] and suggest that the judgment of society as a whole should be embedded into the automated algorithmic systems. However, we lack practical guidance on how to transform understanding of human values and needs, and organizational and societal realities and constraints into the algorithm design process.

We propose a novel method of creating intelligent algorithms — Value-Sensitive Algorithm Design — to address this gap. This method engages stakeholders in early stages of algorithm design and considers human values throughout the design process in a principled and comprehensive manner.

3 METHOD OVERVIEW: VALUE-SENSITIVE ALGORITHM DESIGN

The goal of Value-Sensitive Algorithm Design is to balance multiple stakeholders' values and help achieve collective goals. We begin by considering the notion of "value". We define value as ***"what a person or group of people consider important in life"*** [3]. We interchangeably use the terms motivations, perspectives, needs, and interests.

Our method draws on the principles of user-centered system design [55], value sensitive design [23], and participatory design approaches [51]. Value Sensitive Design (VSD) is a tripartite methodology, consisting of iteratively applied conceptual, empirical, and technical investigations. The goal is to prevent biases in decision-making or compromises in relation to important user values

[3]. The research through design approach suggests that researchers should take "real knowledge" from empirical analysis, "true knowledge" from models and theories of human behavior, and "how knowledge" (the latest technical possibilities) to ideate new systems and technologies and generate reusable knowledge [79]. Our contribution is to transform these design methods created in the context of product or interface design into the context of algorithm design.

The method consists of five steps:

- **Step 1: Understand stakeholders.** The first step is to identify relevant stakeholders (e.g., people who will use the algorithmic tool and people who will be affected by the algorithmic outcomes), and understand stakeholders' motivations, values, and goals.
- **Step 2: Identify algorithmic approaches and create prototypes.** Once the stakeholders have been identified and a thorough investigation of their values has been conducted, designers can identify algorithmic approaches and develop prototype implementations.
- **Step 3: Define methods for working with the community.** The third step is to identify appropriate ways to deploy the algorithmic prototypes, recruit participants, and gather feedback.
- **Step 4: Deploy, refine and iterate the algorithms.** The fourth step is to collect stakeholder feedback and refine the algorithmic prototypes in an interactive iterative process.
- **Step 5: Evaluate algorithms' acceptance, accuracy, and impacts.** The evaluation of the algorithm aims to capture a wide range of factors involved in the real-world multi-stakeholder problem. Specifically, we will evaluate algorithmic tools based on (i) whether they are **acceptable** to the stakeholders' values, (ii) whether they **accurately** solve community problems, and (iii) whether they have positive **impacts** on the community's outcomes and dynamics (we call this AAI evaluation).

Comparison with machine learning-based approach. The most widely-used approaches to develop decision-making or decision-supporting algorithms are driven by Big Data and are machine learning-based. The first step in the process is to define a prediction target. This might consist of whether the defendant will commit a crime if released; whether the child will be removed from the home and placed in care, or whether a job applicant will receive or accept a job offer and be retained for a long time. The second step in the process of developing the algorithm is to use historical data, often in large volumes, for the purpose of training and validating the machine learning models. Finally, the validated models are applied to new data from incoming cases in order to generate predictive scores. However, the single prediction target is often unable to capture the wide range of factors typically involved in any real-world problem. Furthermore, using history to inform the future runs the risk of reinforcing and repeating historical mistakes and fails to capture and incorporate human insights on how the world can be improved in future.

Compared to the automated machine learning-based approach, our method can help to reduce biases in the design choices, increase stakeholders' acceptance and balance stakeholders' values for the following reasons: (1) our method engages stakeholders in the early stages of the algorithm design and uses stakeholders' insights to guide the algorithm creation; (2) our method emphasizes the iterative improving, adapting, and refining based on the stakeholders' feedback; and (3) our method evaluates algorithms not only based on accuracy (which is still important) but also stakeholders' acceptance and impacts of the algorithms on them. However, our method might be more costly because it takes time and effort to gather information from and about stakeholders and involve stakeholders throughout the design process.

"Progress not Perfect". We follow the percept of Value Sensitive Design [23] and the Idea of Justice [65] that *"mere identification of fully just social arrangements is neither necessary nor sufficient"*. The Value-Sensitive Algorithm approach does not aim to eliminate all possible biases and potential

negative impacts from the algorithms. Instead, the goal is to achieve *progress* in improving human well-being.

We will illustrate the method in more detail in the concrete case of designing a recruitment algorithm for WikiProjects in Wikipedia.

4 DESIGNING VALUE-SENSITIVE RECRUITMENT ALGORITHMS FOR PROJECTS IN WIKIPEDIA — A CASE STUDY

4.1 Background

4.1.1 WikiProjects Recruitment. A WikiProject is "a group of contributors who want to work together as a team to improve Wikipedia"[2]. Prior work [21, 41, 78] suggests that WikiProjects provide three valuable support mechanisms for their members: (1) WikiProjects enable members to find help and expert collaborators; (2) WikiProjects can guide members' efforts and explicitly structure members' participation by organizing to-do lists, events like "Collaborations of the Week", and task forces; and (3) WikiProjects can offer new editors "protection" for their work, shielding them from unwarranted reverts and edit wars. As we mentioned in the introduction, despite all the benefits WikiProjects provide, WikiProjects have been suffering decline. Thus, WikiProjects could benefit from infusions of new editors.

However, matching editors to WikiProjects is no trivial task: in the English Wikipedia alone, as of July 2017, there are 500 active WikiProjects covering 3,663,361 encyclopedia articles, 2.9 million active editors, and 38,628 new editors registering on the site in an average month.

The goal of this case study is to **create an intelligent algorithmic tool to match Wikipedia contributors to WikiProjects**.

4.1.2 Wikipedia as Research Platform. Many researchers have sought to conduct experiments and introduce new technologies into online communities like Wikipedia [12, 24, 28, 30, 31, 50, 56, 67]. This is an appealing approach: online communities like Wikipedia constitute a rich laboratory for research: they make collaboration, social interaction, and production processes visible, and offer opportunities for experimental studies. However, to succeed, such studies and experiments require authentic knowledge of the community in question. This is necessary both to (1) create systems that actually solve the intended problems, and (2) conduct work in a way that is acceptable to the community.

However, there has been an unfortunate tradition of academic researchers treating online communities *only* as platforms for their studies, rather than real communities with their own norms, values, and goals. In Wikipedia, research that performs offline analysis of articles and editor actions is typically non-controversial, but studies that involve interventions often encounter resistance from editors, and may sometimes be halted before completion [32]. Various sophisticated algorithmic tools are used to automatically assess the quality of Wikipedia edits and take appropriate actions [24]. Although these tools can efficiently detect and revert low quality edits, research has shown that they also might harm the motivation of well-intentioned new members who are still learning contribution norms [27]. Wikipedia newcomers leave in droves when rudely greeted by algorithmic tools [29]; this violates a community policy — "don't bite newcomers" — and has hindered the growth of the Wikipedia community [66].

Similar problems can occur even for changes to Wikipedia's interface and content introduced by the Wikimedia Foundation. For example, in 2014, the Wikimedia Foundation deployed Media Viewer[3], a change to the MediaWiki software that renders images as an overlay when a reader clicks on them. When this software was first deployed, there was substantial push back from the editing

[2]https://en.wikipedia.org/wiki/Wikipedia:WikiProject

[3]https://www.mediawiki.org/wiki/Extension:Media_Viewer/About

community due to a set of software defects that hid photo licensing information and authorship. The battle that resulted over who has the right to decide what software changes stick (the Wikimedia Foundation or volunteer editor consensus) [69] caused a sudden loss of trust between the developers that maintain the software and the community of editors [54].

The Wikipedia community updated their Wikipedia policy page[4] on *"What Wikipedia is not"* in 2017 to state explicitly that *"research projects that are disruptive to the community or which negatively affect articles — even temporarily — are not allowed and can result in loss of editing privileges"*. Further, the Wikipedia community has developed best practices for *ethically* conducting research on Wikipedia[5], which echo this warning and further suggest that at least one researcher in a study should have "become a [Wikipedia] editor and learned the culture before starting" the study.

To avoid these problems and have our tool succeed in the Wikipedia community, we followed our *Value-Sensitive Algorithm Design* method to create intelligent recruitment algorithms that are sensitive to Wikipedia contributors' interests and motivations and the Wikipedia community's norms and values.

4.2 Step 1. Understand Stakeholders

There are three different types of stakeholders for this research — **newcomers**[6], **WikiProject organizers**, and the **Wikipedia community as a whole**.

To understand the relevant needs and perspectives of newcomers and the Wikipedia community, we rely on extensive bodies of prior work. To summarize, from a newcomer's point of view, participating in Wikipedia and learning the ropes is difficult [27, 43]. WikiProjects are great places to obtain help and receive protection [20]. From Wikipedia's point of view, WikiProjects are an effective mechanism to structure and guide editors' efforts [78] and potentially close the topic coverage gap [40, 74].

However, we know of no prior research that studied the goals and requirements of WikiProject organizers for recruiting new members to their projects. Therefore, we conducted a survey to gather this data to inform our system design. We began by identifying **23** active WikiProjects[7] ranging across a diverse set of topics, including Military history, Medicine, Video games, and Films. We then posted survey questions on the talk pages of these WikiProjects to ask organizers about two key issues: (1) their general interest in recruiting new editors and in using a system that recommends potential new editors for their project; and (2) their current recruitment strategies, including the information they needed to know about potential new members to decide whether to invite them. **59** members from **17** WikiProjects responded to our posts and joined the discussions.

Overall, the survey responses showed that organizers were very interested in recruiting new members, as 29 of them explicitly expressed. For example, one respondent from WikiProject Military History, the largest WikiProject in English Wikipedia, wrote that *"(Recruiting) is very important. We have around 1,000 members and many of them are currently inactive, so we are always looking for new editors to join."* Moreover, project organizers welcomed the prospect of an *intelligent system* to help them in identifying promising candidates to recruit. For example, one respondent wrote that *"We've done invitations manually on the whole, which has a tendency to miss opportunities ... Overall I think this could be a good idea."*

[4]https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not
[5]https://en.wikipedia.org/wiki/Wikipedia:Ethically_researching_Wikipedia
[6]See detailed definition of newcomers in Section 4.3.1
[7]Based on the total number of edits made on the related project pages and project talk pages of each WikiProject.

In sum, the prior work and survey responses collectively suggested **the overall goal of the research — facilitating WikiProject recruitment and encouraging WikiProject participation — is generally aligned with all the stakeholders' interests**.

Furthermore, through synthesizing the prior work and our survey responses, we also identify two potential trade-offs and one key insight (See Table 1 for a summary of trade-offs).

Trade-off 1: "Who to recruit?" Note that the newcomers to WikiProjects might still vary a lot in terms of their experience in Wikipedia. Based on the survey responses, WikiProject organizers tend to prefer editors with some level of experience. For instance, one respondent explicitly wrote that *"It is more important to attract experienced editors than inexperienced new editors."* Another editor wrote *"those who have made several dozen vg (video games)-related edits have already invested in getting acclimated to the arcane laws of WP (Wikipedia) and thus make the best targets for outreach."* However, in Wikipedia, the low retention of new editors has resulted in an overall decline in the active-editor base [27]. Retaining new editors who do not have experience is at least as equally important as retaining experienced editors [29, 52]. Although it is beneficial for experienced editors to join WikiProjects to find collaborators and interesting topics and activities, it is also important (if not more important) for new editors to find mentors and seek assistance in WikiProjects [20].

Take-off 2: "How to determine the fit?" Prior literature [59] discussed two reasons why newcomers choose to participate in a specific online group like WikiProjects: (1) they identify with the group's purpose (interest-based), or (2) they have personal connections with current members of the community (relationship-based). Empirical research [77] analyzed 15-year historical data of over one thousand WikiProjects and confirmed that newcomers with high interest-based connections and the ones with high relationship-based connections both thrive in the WikiProject. However, some WikiProject organizers tend to believe that the alignment in edit interest indicates better fit. For example, one organizer wrote: *"(We want to) invite anyone who has a couple of hundred edits in the Milhist (Military history) area and has edited in the last two weeks."* One respondent wrote explicitly: *"Interaction with other Milhist editors not enough to indicate interest in Milhist"*.

Key insight: "Organizers in the loop". We could have automated the process of inviting new members to WikiProjects, automatically sending invitation messages and even suggesting experienced project members as mentors. However, project organizers wanted to *"be in the loop"* and to maintain the key role in inviting newcomers. One respondent wrote: *"The point is to help editors be better community members, not to coldly spam invites by algorithm."* Survey respondents also provided us some insights on how to communicate the results of the algorithms to WikiProject organizers, as many of them pointed out that they wanted not just a list of recommendations, but also these editors' editing information (such as editing patterns and activity level) and the logic behind the recommendation. For example, one respondent wrote that she/he wants to see a table listing editors' attributes: *"The table could help us identify editors with a propensity for staying, who have already showed internal motivation in learning more about WP but could perhaps use some extra support."*

4.3 Step 2. Identify Algorithmic Approaches and Create Prototypes

We adapt the value analytic method [48] to balance different stakeholders' values when we design algorithm prototypes. First, we will remove design options that some stakeholders strongly object to, such as completely automating the process. Second, design options that some stakeholders find appealing (and others do not strongly object to) will be foregrounded in the design. Third, we will adopt a parallel prototyping approach – implementing and deploying multiple designs to handle varying preferences among stakeholders. We apply these principles to our algorithm designs.

Design Components		Stakeholders in Wikipedia		
		Newcomers	WikiProject Organizers	Wikipedia as a Whole
Who to Recruit	Experienced Newcomers	(+) find collaborators (+) find related topics	(+) quickly adapt and contribute	(+) retain active editors (+) increase overall contributions
	Inexperienced Newcomers	(+) find help to start with (+) find mentors	(-) not the most productive editors to recruit	(+) help alleviate the decreasing trend of the active-editor base
How to Determine the Fit	Interest-based Match	(+) find interested topics	(+) identify reliable contributors	
	Relationship-based Match	(+) find familiar collaborators	(-) maybe not the most productive editors	

Table 1. Perspectives of different stakeholders regarding new member recruitment in WikiProjects.

4.3.1 Candidate selection. Based on the results of the previous step, we created algorithm prototypes for recommending potential new members to WikiProject organizers. However, we first had to define the set of Wikipedia editors for the algorithms to evaluate:

- *Editors had to be active*; they had to have made five or more edits in the previous month [22, 50].
- *Editors should not already be involved with a project*; anyone who already had edited the *project page* for a WikiProject would not be recommended. This indicated they already had knowledge of WikiProjects.
- *Editors could be new to Wikipedia or moderately experienced (but not highly experienced).* This instantiated the trade-off we identified previously. Socializing and retaining "brand new" editors is very important for Wikipedia, but it is more difficult to predict how well-suited they are for a project or how good a job they will do after joining the project. Therefore, while we decided to include these new-to-Wikipedia editors (whom we refer to as "inexperienced newcomers"), we also included more experienced editors who were not yet involved in a WikiProject (whom we refer to as "experienced newcomers"). We operationalized them following the Wikimedia Foundation's guidelines [22] :
  - ***Inexperienced newcomers*** are editors who have successfully completed at least five but less than one hundred edits.
  - ***Experienced newcomers*** are editors who have successfully completed at least one hundred but less than one thousand edits.

We ruled out editors who had made more than one thousand edits in Wikipedia, as they tend to be well socialized into Wikipedia, have their own editing routines, and are likely already to be aware of WikiProjects.

4.3.2 Recommendation algorithms. To balance the different perspectives on how the fit should be defined, we created four different recommendation algorithms and grouped them into two general approaches.

- *Interest-based* algorithms rank candidate editors based on how closely their editing history matches the topic of a WikiProject.
  - The **rule-based** algorithm ranks the match of an editor to a WikiProject by counting the number of (recent) edits by that editor to articles within the scope of the project. Such edits are a strong indicator that the editor is interested in the project's topic.
  - The **category-based** algorithm ranks editors by computing a similarity score between an editor's edit history and the topic of a WikiProject. We followed prior research [8, 77] to represent both editors' histories and WikiProjects' topics as vectors whose elements represent each of the 12 top-level Wikipedia categories[8]. Each element is a real number

[8]https://en.wikipedia.org/wiki/Portal:Contents/Categories

between 0 and 1 representing the degree of interest in the corresponding category. Editors' vectors are calculated based on the Wikipedia categories of articles they edited, and projects' vectors are aggregated from the categories of all articles within their scope. We use the standard cosine similarity metric to compute similarity between the two vectors.

- *Relationship-based* algorithms rank candidate editors based on relationships with current members of a WikiProject.
  - The **bonds-based** algorithm ranks editors by the strength of "social connections" the editor has to current members of a WikiProject. Prior research [58, 77] suggests that such social connections are a good predictor of a new editor becoming a successful contributor to a WikiProject. We followed this prior research to operationalize social connections by counting the number of edits a candidate editor made to the user talk pages of current members of a WikiProject.
  - The **co-edit-based** algorithm is a version of collaborative filtering [60, 63] and is inspired by the design of SuggestBot [12]. Candidate editors are ranked by the similarity of their edit histories to the edit histories of current members of a WikiProject. We use the approach of Warncke-Wang et al. [76] to handle the sparse overlap of edits between two editors. Specifically, the similarity between two editors is calculated as the intersection of articles the two editors edited divided by the multiplication of the square roots of the numbers of unique articles those two editors edited.

Note that if we rank experienced newcomers and inexperienced newcomers all together, the experienced ones might overshadow the inexperienced ones. Therefore, we decided to rank the two types of newcomers separately.

4.3.3 Presentation and explanation of the recommendations. We also designed a user interface for presenting recommended new editors to WikiProject organizers. We decided to include several recommendations from each of the algorithms, resulting in a total of about 12 recommended new editors per batch. The user interface was implemented as an interactive element within a Wikipedia page. Recommendations were presented in a sortable table that showed basic information about each candidate editor, such as their registration date and the total number of Wikipedia edits. We also provided explanations for each recommendation; the explanations were based on the research that motivated each algorithm. To take one example, the explanation for the bonds-based algorithm followed the findings of [77] to suggest that these editors were likely to have good retention within the project. Figure 1 shows what our user interface[9] looks like after deployment and refinement in response to organizers' feedback.

4.4 Step 3: Define methods for working with the community

We took several approaches to ensure that *how we conducted our research* was consistent with Wikipedia norms and acceptable to the WikiProject communities.

- ***We communicated our research plan early with the community***. We published an initial research plan on the Wikimedia Meta-Wiki[10], a global community forum for Wikimedia projects and activities. We also shared the plan through other relevant discussion boards and public channels, including Wikimedia research community mailing lists[11], WikiProject

[9] This is populated with made up data to preserve participant privacy.
[10] https://meta.wikimedia.org/wiki/Main_Page
[11] https://lists.wikimedia.org/mailman/listinfo/wikimedia-l

Username	Why we recommend this editor	First Edit Date	Total Edits in ENWP	Editor Status	Invite	Survey
Hulk	Hulk made 77 out of their total 187 edits to articles within the scope of your project.	2017-8-24	187	Experienced Editor	invite	survey
Wolverine	Wolverine made 49 out of their total 87 edits to articles within the scope of your project.	2017-7-23	87	Newcomer	invite	survey
Iron Man	Iron Man edited articles similar to articles your project members edited. For example, Iron Man and you project member Superman edited 9 same articles in their most recent 200 edits.	2017-7-26	191	Experienced Editor	invite	survey
Thor	Thor edited articles similar to articles your project members edited. For example, Thor and you project member Captain America edited 8 same articles in their most recent 200 edits.	2017-5-19	52	Newcomer	invite	survey

Fig. 1. Prototype newcomer recommendations for WikiProject organizers. We embed the explanation of why the editor can be a useful contributor and the definition of editor status in the rollover text.

X[12], the Wikipedia Village pump[13], and the Wikipedia Teahouse[14]. This let us obtain and incorporate input from relevant stakeholders, and also helped inform several key system and study design decisions.

- ***Recruit participants using methods consistent with the community's norms and practices.*** We recruited WikiProject organizers through an opt-in process; the lead researcher created a sign-up table on the talk page of his Wikipedia account for organizers to register their interest. To bring this to the community's attention, we collaborated with several active participants who publicized our study. We also worked with the Wikipedia Signpost[15] to write an article explaining our study and informing WikiProject organizers how to participate.

4.5 Step 4: Deploy, Refine, and Iterate

As noted above, we delivered four distinct batches of recommended new editors to WikiProject organizers over a 6-month period. Each batch included a brief survey that asked the organizers to rate the quality of the recommendations and give overall feedback. We made refinements to our system after Batches 1 and 2. We describe the changes we made the system in this subsection.

4.5.1 From Batch 1 to Batch 2. Based on the survey responses and other feedback, we made the following changes to our system before computing and delivering the second batch of recommendations.

- ***Refine algorithm explanations.*** Our initial research-based explanations were not helpful to project organizers. Instead, we rewrote them to focus on why recommended new editors might be useful contributors to their projects.
- ***Add filters to screen out candidate editors who have been blocked, banned, or received warning messages.*** Our initial algorithms could recommend any sufficiently active editors. However, WikiProject organizers noticed that we recommended some editors with problematic edit histories, e.g, some of them had been blocked, banned, or warned for vandalism, or had been reverted frequently. As one organizer said:
  "[The] user's edits were immediately reverted, looked like they started out trying to be helpful but became outright vandalism."
- ***Temporarily remove bonds-based and category-based matching algorithms***, because some project organizers were confused by the rationales behind these algorithms.

[12]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_X
[13]https://en.wikipedia.org/wiki/Wikipedia:Village_pump
[14]https://en.wikipedia.org/wiki/Wikipedia:Teahouse/About
[15]https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost

4.5.2 From Batch 2 to Batches 3 & 4. At this point, organizers generally were satisfied. However, we identified several changes to enhance the system effectiveness.

- ***Boost thresholds in matching algorithms*** to improve recommendation quality. For example, candidate editors will be recommended by the rule-based matching algorithm only if they made at least five edits on project-related articles in the previous month (instead of any edits, as in the original version).
- ***Reinstate the category-based matching algorithm***, and update the explanation.

4.6 Step 5: Evaluate Algorithms' Acceptance, Accuracy, and Impacts

In the next section, we evaluate our algorithmic system based on (1) whether it is *Acceptable* to the Wikipedia community's values, (2) whether it *Accurately* solves community problems, and (3) whether it has positive *Impacts* on the community's outcomes and processes.

4.6.1 Community Acceptance
. The survey responses and other feedback from organizers and new editors were *positive*. This gave us confidence that our system was *acceptable* to the community. One project organizer wrote:

> *"I really think [Anonymized Researcher] is on to something here. To date, recruitment of new editors to WikiProject Military history has been rather organic, with coordinators inviting them to join when we notice someone new via our watchlisted articles. But no-one has all our articles watchlisted. This puts some science behind recommendations, and will be a great supplement to the current processes. We've already had new members join thanks to [Anonymized Researcher]'s tool. I'm looking forward to seeing where this goes."*

Another organizer wrote:

> *"My experience has been positive. The majority of the recommendations have been clueful and we've had a few positive responses. "*

An organizer pointed out a key value of our system to them:

> *"I would recommend the tool to others as a valuable way to reach new editors who might otherwise slip through the cracks, and who might prove to be valuable members of the community. Early guidance in Wikipedia's rules, customs, and culture is likely to make for a smooth transition from new editor to established editor, and I believe this tool will enable established users to more effectively and systematically reach a better cross section of new potential editors for the WikiProject."*

Editors who were invited to join projects also reacted positively. For example, an editor invited to join WikiProject Visual arts wrote:

> *"Hi, dear [Organizer], how nice of you to get in touch. I added the sites you mentioned to my bookmarks here. It's a great feeling to be able to connect and ask questions if necessary. Right now I'm on my way to write a new article for the German Wikipedia — and I'm translating the German article on de:Museum Wiesbaden for the English Wikipedia-article, as this has been a stub. But I also added the Cleanup listing for further checking on it. I cannot work many hours a day, but I try to do my best to help! Kind regards, –[New Editor]"*

Another editor invited to WikiProject Africa appreciated the resources the organizer provided and wrote:

> *"Hi [Organizer], This is [New Editor]. Thank you for reaching out to me and thank you for informing me about the WikiProject Africa talk page. I will definitely put that page*

	Algorithm Types				Newcomer Types	
	Rule-based	Category-based	Bonds-based	Co-edit-based	Exp. Newcomers	Inexp. Newcomers
Avg. Ratings	3.24	2.36	2.33	2.76	2.85	2.88
Invitation Rates	47%	16%	22%	28%	34%	32%

Table 2. Average ratings of newcomers of different types and recommended by different algorithms. The ratings are on a 5-point Likert scale.

on my watchlist. The links you sent are interesting, thank you for sharing them with me. Thank you once again for reaching out to me, I appreciate it. [New Editor] "

Given that our system was *acceptable* to the community, we also sought to carefully *evaluate* our system and understand the *impacts* of our system.

4.6.2 *Algorithm Accuracy*

. To evaluate the algorithms, we seek to answer the following three questions.

(1) How do WikiProject organizers value the recommendations delivered by our algorithms?
(2) How do WikiProject organizers act on the recommendations and invite recommended new editors to their projects?
(3) Do any of our algorithms work better than others?

Overall **26** WikiProject organizers from **39** WikiProjects signed up to use the algorithmic tool. **16** organizers representing **18** WikiProjects actively used the tool. During the six-month period, we sent out **385** recommendations. For each recommendation, we included a link that let organizers rate "How good a fit is this editor for your project?" on a 5-point Likert scale, and received **229** ratings in total. Organizers acted on the recommendations and sent out **100** invites to newcomers in total.

Table 2 presents the average rating and invitation rate for the four different types of algorithms, and the average rating for experienced and inexperienced newcomers respectively. Table 3 shows the relationship between ratings and invitations for the two types of newcomers. The results suggest that:

(1) The rule-based algorithm was rated higher and resulted in more invitations compared to the other three types of algorithms: Category-based, Bonds-based and Co-edit-based. The average rating for the rule-based algorithm is 3.24, significantly higher than the other three types of algorithms ($t = 3.51$, $p < .001$). The invitation rate for the rule-based algorithm is 47%, which is also higher than Category-based (16%), Bonds-based (22%), and Co-edit-based (28%). There is no significant difference among the other three algorithms.

(2) The average ratings *(2.85 v.s. 2.88)* and the invitation rates *(34% v.s. 32%)* are not that different between experienced and inexperienced newcomers. Experienced newcomers did not overshadow the inexperienced newcomers — both types of newcomers got an equal chance to be invited to join WikiProjects.

(3) Table 3 shows that newcomers who received low ratings (1 or 2) were rarely invited. This is not surprising. However, not all newcomers who received high ratings (4 or 5) were invited. For example, 35% of the newcomers who received a rating of 5 (i.e., an excellent fit for the project) were not invited. This opens up opportunities for future research to better understand the discrepancy between organizers' ratings and actions, which will help inform algorithm and system designs.

4.6.3 *Impacts on the Community*

. Our evaluation on the impacts on the community seeks to answer the following two questions:

Ratings	1	2	3	4	5	Total
Exp. Newcomers	3%	4%	16%	84%	64%	34%
Inexp. Newcomers	0%	0%	50%	48%	67%	32%
Total	2%	3%	32%	70%	65%	33%

Table 3. Invitation rates for two types of newcomers at different ratings. 5 means the editor is an *"Excellent fit"* for the target WikiProject, and 1 means *"Not a good fit"*.

(1) What happened to the newcomers if WikiProject organizers invite them to the projects? Specifically, do the newcomers participate and contribute? And how is receiving an invitation from an organizer different from receiving template invitations from the researcher team, or receiving nothing?
(2) Are there second-order effects? For example, if invited editors do participate in the projects they were invited to, are they *shifting* their Wikipedia editing there, or *increasing* the overall amount of Wikipedia editing?

We designed our evaluation to systematically answer these questions.

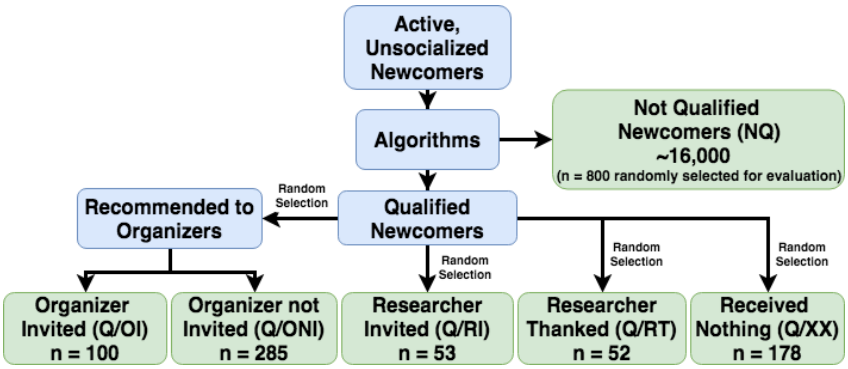


Fig. 2. An overview of how the experimental groups were formed. Green rectangles represent the six groups. Note that *Groups Q/RI, Q/RT, and Q/XX* were only introduced in Batches 3 and 4.

Study Design. Although we made changes after the first two batches, these batches also potentially impacted the community. Therefore, our formal evaluation of the impacts of our system reports on all four batches (Our models include 'batch' as a random effect, thus allowing us to check whether different batches had different impacts).

Our study included six experimental groups, allowing us to evaluate the effects of our algorithms against different relevant baselines (see Figure 2 for an overview). To understand how the groups were formed, we first revisit what we mean by *qualified* editors. Qualified editors are the set of all editors determined to be suitable new members for any of the WikiProjects in our study by any of our algorithms. Thus, *unqualified* editors are not necessarily "bad Wikipedia editors". They simply were not assessed to be suitable for any of the target WikiProjects by algorithms.

- ***Not Qualified (NQ)***. Active editors that none of our algorithms judged to be relevant to any of our target WikiProjects. We randomly selected 800 editors (200 per recommendation batch) for this evaluation purpose.

- ***Qualified, Received Nothing (Q/XX)***. Qualified editors who received no treatment; they were not presented to organizers, and they received no message. This group serves two purposes: (1) By comparing *"NQ"* and *"Q/XX"*, we can test whether our algorithms actually distinguish "qualified" and "unqualified" newcomers. (2) By comparing this group to all the other groups that received a treatment, we can evaluate the effects of those treatments.
- ***Qualified, Organizer Invited (Q/OI)***. These qualified editors were included in the recommendations delivered to WikiProject organizers and were invited by the organizers. We note that WikiProject organizers may send a one-click invitation using our template message or customize the invitation message before sending. The majority of organizers chose to customize their invitations.
- ***Qualified, Organizer not Invited (Q/ONI)***. These qualified editors were included in the recommendations delivered to WikiProject organizers but were not invited by the organizers.
- ***Qualified, Researcher Invited (Q/RI)***. We were interested in evaluating the efficacy of automated invitations to new editors. Therefore, we randomly selected a subset of qualified editors not included in the recommendations sent to project organizers, and we sent them the general template invitation message (from the lead researcher's Wikipedia account).
- ***Qualified, Researcher Thanked (Q/RT)***. The research literature shows that a simple socialization message or even just an acknowledgement can have a significant positive influence on new editors [9]. Therefore, we randomly selected a subset of qualified editors not included in the recommendations sent to project organizers, and sent them a simple "thank you for your effort" message (again from the lead researcher's Wikipedia account).

Outcome variables. We measured the effect of our system on new editors, specifically in terms of the number of their edits within the relevant WikiProjects and in the rest of Wikipedia[16]. Thus, our outcome variables were:

- ***Within-project edits***: We define the edits the editor made on any article claimed *within the scope* of the target WikiProject (including edits on both article pages and article talk pages) as within-project edits[17]. We computed the change of edits the editor made one week before and one week after[18] the point of algorithm evaluation (for groups that do not involve interventions — *Groups: NQ, Q/XX, and Q/ONI*) or the point of receiving the intervention message (for groups that involve interventions — *Groups: Q/OI, Q/RI, and Q/RT*).
- ***Outside-project edits***: The metric is computed the same as within-project edits except that we calculated the number edits made outside the target project's scope (e.g., the total edits made in Wikipedia subtracted by with-project edits).

Statistical Model. We used a linear random-effect (mixed) model with each batch as a group[19]. We used the model to examine how the six different editor groups (i.e., *Q/OI, Q/ONI, Q/RI, Q/RT, Q/XX*, and *NQ*) performed differently in terms of their within-project and outside-project contributions after being evaluated or invited.

We used condition *Q/XX* as the baseline in the model for the reasons we stated earlier: (1) by comparing *NQ* and *Q/XX*, we can test whether our algorithms actually distinguish "qualified" and

[16] We explored two additional outcome variables in our exploratory analysis: (1) Withdrawal from WikProjects and Wikipedia (with survival analysis), and we draw very similar conclusions compared to the edits-based outcomes; (2) Quality change before and after the interventions, and we did not see any significant change over the observational window.

[17] For newcomers in *Group: NQ*, we considered all the WikiProjects signed up in our study as their potential target WikiProjects, and computed the average number as their within-project edits.

[18] We tried both one week and one month as the observation window for our analysis, which showed very similar patterns. Therefore, in the paper we only report the results with one week as the observation window.

[19] Stata command: *xtreg, re.*

	Experienced newcomers						Inexperienced newcomers					
	Within project Model 1			Outside project Model 2			Within project Model 3			Outside project Model 4		
	Coef.	S.E.	95% C.I.	Coef.	S.E.	95% C.I.	Coef.	S.E.	95% C.I.	Coef.	S.E.	95% C.I.
Q/XX (Intercept)	.38 ***	.09	[.21, .56]	.94 *	.44	[.08, 1.81]	.18 **	.06	[.06, .29]	.91 ***	.22	[.46, 1.34]
NQ v.s. Q/XX	-.36 ***	.10	[-.56, -.16]	.61	.50	[-.37, .1.60]	-.17 **	.06	[-.29, .05]	-.45	.24	[-.92, .02]
Q/OI v.s. Q/XX	.43 **	.16	[.13, .74]	.54	.77	[-.97, 2.04]	.01	.09	[-.16, .19]	-.44	.35	[-1.12, .24]
Q/ONI v.s. Q/XX	-.05	.12	[-.28, .19]	.03	.58	[-1.11, 1.17]	-.09	.07	[-.24, .05]	-.47	.29	[-1.03, .09]
Q/RI v.s. Q/XX	-.05	.21	[-.46, .36]	.17	1.04	[-1.86, 2.20]	-.06	.10	[-.27, .14]	-.57	.41	[-1.37, .23]
Q/RT v.s. Q/XX	-.34	.19	[-.70, .03]	.62	.92	[-1.18, 2.42]	.16	.10	[-.04, .36]	-.70	.40	[-1.48, .09]

Table 4. Summary of the change of new editors' activities in different groups using *Group: Q/XX* as the model baseline. *** p <.001, ** p <.01, * p <.05. Notice that *Groups Q/XX, Q/RI, and Q/RT* are only introduced in Batches 3 and 4 and *Groups NQ, Q/OI, and Q/ONI* span all the batches, therefore we used the batch number as random effects in the models to eliminate its effects and to ensure the correctness of our results.

"unqualified" newcomers. (2) By comparing *Q/XX* to all the groups that received a treatment, we can evaluate the effects of those treatments.

Consider model 1 as an example. The predicted within-project edits for the baseline *Group: Q/XX* is 0.38 (intercept). Because we operationalize within-project edits as the edits change over the observation period (in ratio), we interpret the result as editors' edits in *Group: Q/XX* on average dropped 1 - (.38) = 62% over the observation period. In model 1, the coefficient of *NQ* v.s. *Q/XX* is -0.36, which suggests that the predicted value for *Group: NQ* is 0.02. Therefore, we interpret the result as editors' edits in *Group: NQ* on average dropped 1 - (.02) = 98%. We interpreted the results in the same way in the rest of the paper.

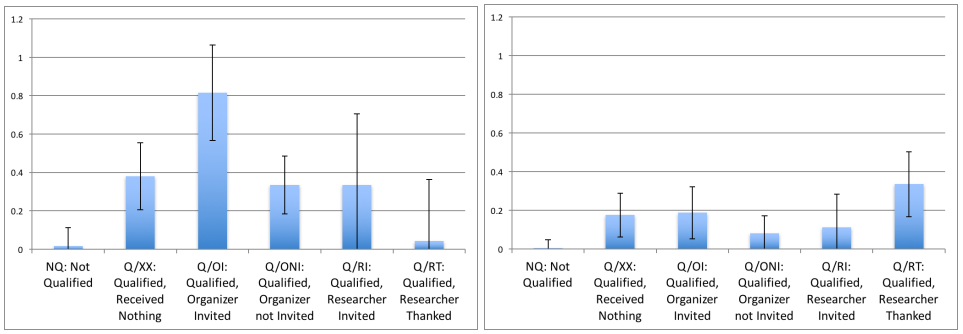
Results. We now present our answers to our questions.

Question 1: What happened to the newcomers if they received invitations from WikiProject organizers? Specifically, do the newcomers participate and contribute? And how is receiving organizers' invitations different from receiving template invitations from the researcher team, or receiving nothing?

In Models 1 and 3 (in Table 4 and Figure 3), we compared the within-project contributions of the six newcomer groups (*i.e., NQ, Q/XX, Q/OI, Q/ONI, Q/RI, and Q/RT*), separating *experienced* and *inexperienced* newcomers. We used *Group: Q/XX, i.e., Qualified, Received Nothing* as the baseline in our models. The results of our models are:

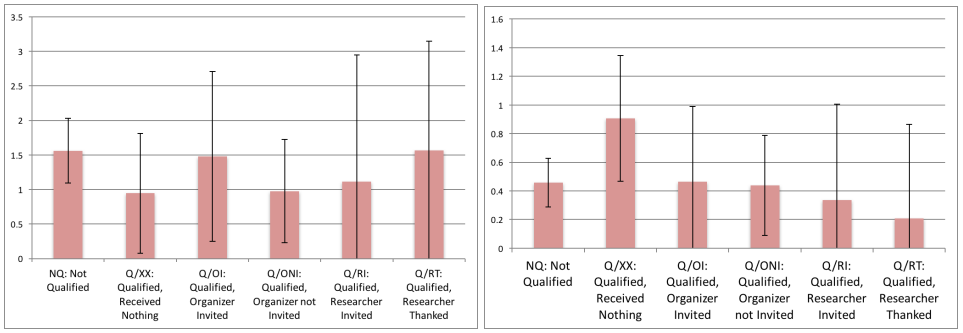
(1) WikiProject newcomers' (both unqualified and qualified newcomers) in-project contributions on average dropped sharply over the observation period. This pattern is consistent with the findings of prior work [9] and confirms the challenge of newcomer socialization in Wikipedia [43, 68]. *However, qualified newcomers dropped less than unqualified newcomers.* On average unqualified experienced newcomers (*NQ*) decreased their edits by 98% in the two week observation period, while qualified experienced newcomers who received nothing (*Q/XX*) decreased by 62%. The difference is statistically significant (Coef. = -0.36, p <.001). Unqualified inexperienced newcomers (*NQ*) decreased their edits by 99% in the two weeks' observation period, while qualified experienced newcomers who received nothing (*Q/XX*) decreased by 82%. The difference is statistically significant (Coef. = 0.17, p <.001). This provides evidence that our algorithms can identify newcomers who are more suitable and have potential to become serious contributors for the target WikiProjects.

(2) Experienced newcomers who received an invitation from a WikiProject organizer (*Q/OI*) had a 43% increase in their in-project activity over the baseline *Group: Q/XX*. But we do not see the same pattern among inexperienced newcomers. Experienced newcomers in the baseline



(a) Within-project edits: exp. newcomers (b) Within-project edits: inexp. newcomers

Fig. 3. The effects of algorithm-based interventions on newcomers' within-project contributions. The plots show the predicted values, i.e., value 1.0 at the y-axis indicates editor's contributions did not change during the observation period. The error bar indicates the 95% confidence interval.



(a) Outside-project edits: exp. newcomers (b) Outside-project edits: inexp. newcomers

Fig. 4. The effects of algorithm-based interventions on newcomers' outside-project contributions. The plots show the predicted values, i.e., value 1.0 at the y-axis indicates editor's contributions did not change during the observation period. The error bar indicates the 95% confidence interval.

(*Group: Q/XX*) reduced their edits by 62% over the two-week observation period. In contrast, experienced newcomers in *Group: Q/OI* only decreased by 19%. The difference is statistically significant (Coef. = .43, p <.01). However, we do not see the same pattern among inexperienced newcomers. Inexperienced newcomers who received invitations from WikiProject organizers were statistically indistinguishable from the baseline group (Coef. = .01, p n.s.).

(3) Receiving an invitation or thank-you message from the researcher had no impact on newcomers' within-project contributions, whether they were experienced or inexperienced. There were no significant differences between the baseline (*Q/XX*), newcomers who received invitations from the researcher (*Q/RI*), and newcomers who received thank-you messages (*Q/RT*). This was true for both experienced newcomers (Coef. *Q/RI* v.s. *Q/XX* = -0.05, p n.s.; Coef. *Q/RT* v.s. *Q/XX* = -0.34, p n.s.) and inexperienced newcomers (Coef. *Q/RI* v.s. *Q/XX* = -0.06, p n.s.; *Coef. Q/RT* v.s. *Q/XX* = 0.16, p n.s.).

Question 2. Are there second-order effects? The key point here is to determine whether *increased activity within a WikiProject* led to *reduced activity outside the project*; if so, this suggests the effect of the intervention is to *redirect* a finite amount of editor effort, rather than *increase*

overall editor effort. Even the former can be useful, if effort is redirected to areas of community need [35, 44, 75]. However, from the point of Wikipedia as a whole, the latter obviously is preferable.

Model 2 (shown in Table 4 and Figure 4) shows that the increase in edits within target WikiProjects by experienced newcomers in *Group: Q/OI* (those who received an invitation from WikiProject organizers) did not come at the cost of fewer edits in the rest of Wikipedia. Moreover, none of the groups that received messages (*Groups: Q/OI, Q/RI, and Q/RT*) were statistically different from the baseline (*Group: Q/XX*) in outside-project contributions for either experienced or inexperienced newcomers. Therefore, we found no evidence that the system treatments negatively impacted outside-project contributions.

5 DISCUSSION

5.1 Summary of the case study

The case study of WikiProject intelligent recruitment tool illustrated the Value-Sensitive Algorithm Design method. Our evaluation showed that (1) the algorithmic tool was well received by the community; (2) inexperienced editors and experienced editors were equally invited by the project organizers through the recruitment algorithms; (3) only experienced editors who received invitations from project organizers had a significant increase in their within-project activities over the baseline group; and (4) the increase in the invited experienced editors' contributions did not come at the cost of fewer edits in the rest of Wikipedia.

One question we consider is: why did only one experimental group — qualified experienced newcomers who received invitations from project organizers — see an increase in their activity over the baseline? We consider two factors that could play a role in this outcome: (1) receiving a message from organizers; (2) being experienced rather than inexperienced.

Regarding *why receiving a message from project organizers* has had a positive impact, we consider four possibilities.

(i) It wasn't receiving the message, it was being selected by the organizer. The simplest explanation is that project organizers did additional investigation of candidate editors before inviting them. Our interface gave organizers one-click access to candidate editors' user talk pages, from which they could learn about their edit history. Therefore, perhaps organizers only selected the most promising candidate editors.

(ii) It was the interaction between the organizer and the new editor. Once organizers invited a new editor to their project, they also could interact with and mentor that new editor, which generally would have a positive impact on the new editor's activity within the project. We saw evidence to support this possibility. Here is what a new editor told us: *"I've received outreach from other member in the project, and I've been helped out ... When you join a project like Chicago it's so much work to be done. But I've received help from [Anonymized name] the organizer so I don't feel like I am doing it alone. I feel that's really helpful."* We also received some inquiries from new editors in response to the automated invitations we sent. However, it was not in our protocol to respond, nor would we have had the domain knowledge to do so in any case.

(iii) It was the organizer's record and reputation. Perhaps it was motivating to be invited by a project organizer with an established record. For example, one invited new editor wrote: *"I would be honored to work with [Organizer] on any articles he needs help with."*

(iv) It was the wording of the organizer's invitation. Since project organizers usually customized the invitation messages they sent out, perhaps those messages were tailored to the invited editors, and thus were more motivating.

And of course, some combination of all these factors may be having an effect. Therefore, future work is needed to tease apart the effects. For example, we might ask project organizers to identify

all the candidate new editors that they *would choose to invite*. We then could control the treatments these editors receive, such as a personalized invitation from the organizer, an automated template invitation, an automated thank-you, or nothing at all.

Regarding *why being invited to a project by an organizer had no effect on inexperienced newcomers*, the most plausible explanation is that these editors were still too new to Wikipedia. These editors may still have so much to learn about Wikipedia skills and norms that they cannot take advantage of an invitation from a project organizer. There was some evidence for the this possibility, as seen in several messages from inexperienced newcomers to the organizers who invited them: *"Hi, I received an invitation from you to join this project following edits I'd apparently made relating to it ... I can't find the project page you refer to, so I can't really help, sorry!"* In future work, we will explore how to better support the inexperienced newcomers with potential. One possible way is to first direct inexperienced editors to useful resources such as Teahouse [50] in order to help them become accustomed to Wikipedia culture, ask questions, and develop community relationships, instead of asking the inexperienced newcomers to contribute to projects right away.

5.2 Reflection on the Value-Sensitive Algorithm Design Method

Twenty-one years ago, Ben Shneiderman and Pattie Maes had a famous debate on direct manipulation versus automated agency on user interface design at CHI'97 [46]. "Direct manipulation" suggests enhancing a user's ability to directly manipulate interfaces, access information, and invoke services, while "automated agency" centers on building machinery for sensing a user's activity and taking automated actions. In 1999, Eric Horvitz [33] proposed twelve "Mixed-Initiative" principles that provides a foundation for integrating research on direct manipulation and automated agency. The goal of mixed-initiative design is to allow intelligent services and users to collaborate efficiently to achieve a user's goals.

Similar to the mixed-initiative approach in user interface design, the goal of Value-Sensitive Algorithm Design is to integrate **human control** and **automated systems**. We avoid building complex systems that rely solely on automatic data mining. On the other hand, we do not want to revert to complete human control and ignore the significant efficiencies gained from automated approaches. Value-sensitive algorithm design attempts to seek valuable synergies between human control and automated systems, especially through:

- **Using multiple relevant stakeholders' insights to guide the specific algorithmic choices.** Relating to the case study, we translated newcomers', WikiProject organizers', and the Wikipedia community's "values" into algorithm design choices, including the definition of the inclusion criteria and four different matching algorithms — the rule-based algorithm, the category-based matching algorithm, the bond-based algorithm, and the co-edit-based algorithm.
- **Considering social and organizational context in the algorithm creation.** Specifically, we want to design algorithms that can facilitate, augment, and improve current organizational processes. The idea is similar to the principle of "value-added automation" proposed in [33]. Relating to our case study, based on the suggestions of community stakeholders, we decided that current project organizers should have the final say on who should be invited and manually extend invitations, rather than automating the whole invitation process.
- **Working closely with relevant stakeholders and engaging them in the early iteration and refinements.** In our case study, we communicated our research plan early with the community and proactively initiated discussions on relevant discussion boards and public channels. We iteratively pre-tested and refined our algorithm design with a small set of WikiProjects before deploying the tool at large scale.

- **Evaluating not only the accuracy, but also the acceptability and impacts of algorithms.** In our case study, we evaluated our algorithmic tool based on (1) whether it was acceptable to project organizers and new members, (2) whether it accurately solved community problems by measuring the organizers' subjective ratings of the algorithm output and the invitation rate, and (3) whether the algorithm had a positive or negative impact on the invited newcomers' subsequent performance.

The Value-Sensitive Algorithm Design method suggests a specific 5-step procedure on how to engage in a particular kind of algorithm design inquiry. However, like other design methods, "the execution of a method may correspond more or less closely to its descriptive form" [23]. The value-sensitive algorithm design method is intended to be open to adaption and evolution, in response to the actual design situation.

5.3 Applications of the Method

We see the Value-Sensitive Algorithm Design method best suited for solving "wicked" problems [4] that lack clear ground truth. Recruitment algorithms, as well as the examples mentioned in the introduction, such as recidivism risk assessment algorithms and risk modeling algorithms for the maltreatment of children, are examples. For these problems, the actual evaluation outcomes (e.g., the suitability of candidates to an group or an organization, the risk of re-offending, and the severity of child maltreatment) are difficult to observe. However, future studies need to be conducted to verify whether this method can be applied in other contexts.

5.4 Challenges and Opportunities

Our case study about creating recruitment algorithms for WikiProjects demonstrated the value and promise of the Value-Sensitive Algorithm Design method. However, it also surfaced difficult challenges. Specifically, it is challenging to explain and articulate our algorithms in ways that enable stakeholders to understand them sufficiently and provide sensible feedback to improve them; this constitutes a barrier for stakeholders to be fully engaged in the process. We note that researchers in Explainable Artificial Intelligence (XAI) have made great progress on transforming complex models, such as neural networks, into simple ones, through approximation of the entire model [13, 70] or local approximation [61]. Despite the mathematical rigor, there are recent critiques that this line of research is built based on the intuition of researchers, not on a deep understanding of actual users [49]. That is, there is limited empirical evidence whether these intelligible models and explanation interfaces are actually understandable, usable, and practical in real-world situations [1, 17]. Therefore, one promising research direction is to improve the usability of explanation interfaces for non-expert users in real-world situations and perform empirical studies to evaluate the efficacy of these interfaces.

The second challenge is how to address the fundamental mismatch between human styles of interpretation, reasoning, and inputs and statistical optimizations of high-dimensional data [5]. Specifically, how can we effectively translate human insights and inputs to adjust and improve algorithms? There is lack of holistic understanding of the mapping space between human inputs and algorithmic design options. Furthermore, there is no interface support to effectively elicit users' insights about how to improve algorithms and visualize their influence.

The third challenge is that if we allow a massive number of people with diverse values, preferences, opinions, and interests to influence algorithms, how do we appropriately aggregate their inputs to ensure algorithm justice and accountability? For example, how can we deal with value conflicts? How can deal with stakeholders' different levels of participation at different design stages? Note that many intelligent algorithms (including the recruitment algorithm in our case study, recidivism

prediction, and child maltreatment prediction algorithms mentioned in the introduction) are designed to address a "collective problem" that by nature will impact large groups of people with diverse perspectives and interests simultaneously. In other words, individualized and personalized approaches of adjusting algorithms are not sufficient. It would be interesting to explore socio-technical solutions to allow people to collectively design, train and modify the algorithms and publicly negotiate the outcomes.

Another challenge is that companies or government agencies might intentionally hide algorithmic decision-making procedures from public scrutiny. This challenge has to be addressed through legislative efforts.

To bridge the gap between algorithm experts and non-expert stakeholders, educational innovation is also critical. In our vision of algorithm education, on the one hand, we need to teach future algorithm developers to think not just about "solving problems" but also to respect the needs, motivations, and values of stakeholders. Human well-being should be considered throughout the process of algorithm development in a principled and comprehensive manner. On the other hand, we want to educate an algorithmically literate society and promote "algorithmic thinking" at all levels. For example, we might want to teach students how to reason about application/technological advances, algorithm concepts, and intelligent systems, and how to understand such systems.

6 CONCLUSION

We proposed a novel approach to the design of algorithms, which we call Value-Sensitive Algorithm Design. This approach incorporates stakeholders' tacit knowledge and insights throughout the design process. As a case study, we designed and deployed an algorithmic tool to help WikiProjects identify and recruit new members; this study demonstrates the value and promise of the new method and surfaces important new challenges.

7 ACKNOWLEDGMENT

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[2] Julia Angwin. 2016. Make algorithms accountable. *The New York Times* 1 (2016), 168.

[3] Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1125–1134.

[4] Richard Buchanan. 1992. Wicked problems in design thinking. *Design issues* 8, 2 (1992), 5–21.

[5] Jenna Burrell. 2016. How the machine âĂŸthinksâĂŹ: Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[6] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. Controlling attribute effect in linear regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 71–80.

[7] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[8] Jilin Chen, Yuqing Ren, and John Riedl. 2010. The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 821–830.

[9] Boreum Choi, Kira Alexander, Robert E Kraut, and John M Levine. 2010. Socialization tactics in wikipedia and their effects. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 107–116.

[10] 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 *Conference on Fairness, Accountability and Transparency*. 134–148.

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[12] 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*. ACM, 32–41.
[13] Mark Craven and Jude Shavlik. 1999. Rule extraction: Where do we go from here. *University of Wisconsin Machine Learning Research Group working Paper* 99 (1999).
[14] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.
[15] 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.
[16] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2016. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2016), 1155–1170.
[17] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
[18] Benjamin G Edelman and Michael Luca. 2014. Digital discrimination: The case of airbnb. com. (2014).
[19] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.
[20] Andrea Forte, Niki Kittur, Vanessa Larco, Haiyi Zhu, Amy Bruckman, and Robert E Kraut. 2012. Coordination and beyond: social functions of groups in open content production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 417–426.
[21] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26, 1 (2009), 49–72.
[22] Wikimedia Foundation. [n. d.]. Definition of research terms in Wikimedia Foundation. ([n. d.]). http://upload.wikimedia.org/wikipedia/commons/0/09/Definitions_of_Research_Terms.pdf
[23] Batya Friedman, David G. Hendry, and Alan Borning. 2017. A Survey of Value Sensitive Design Methods. *Foundations and TrendsÂő in HumanâĂŞComputer Interaction* 11, 2 (2017), 63–125. https://doi.org/10.1561/1100000015
[24] 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*. ACM, 6.
[25] Sara Hajian and Josep Domingo-Ferrer. 2013. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* 25, 7 (2013), 1445–1459.
[26] Sara Hajian, Josep Domingo-Ferrer, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2015. Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1733–1782.
[27] 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.
[28] Aaron Halfaker, R Stuart Geiger, and Loren G Terveen. 2014. Snuggle: Designing for efficient socialization and ideological critique. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 311–320.
[29] 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*. ACM, 163–172.
[30] Aaron Halfaker, Bryan Song, D Alex Stuart, Aniket Kittur, and John Riedl. 2011. NICE: Social translucence through UI intervention. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM, 101–104.
[31] Benjamin V Hanrahan, Gregorio Convertino, and Les Nelson. 2012. Modeling problem difficulty and expertise in stackoverflow. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*. ACM, 91–94.
[32] Marit Hinnosaar, Toomas Hinnosaar, Michael E Kummer, and Olga Slivko. 2017. Wikipedia Matters. (2017).
[33] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 159–166.
[34] Andrew McAfee Natalie Saltiel Iyad Rahwan, Karthik Dinakar. 2016. Society in the Loop Artificial Intelligence. (2016). https://joi.ito.com/weblog/2016/06/23/society-in-the-.html
[35] Isaac L Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at home on the range: Peer production and the urban/rural divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 13–25.
[36] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*. IEEE, 1–6.
[37] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 869–874.

[38] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems* 35, 3 (2013), 613–644.
[39] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.
[40] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2009. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1509–1512.
[41] Aniket Kittur, Bryant Lee, and Robert E Kraut. 2009. Coordination in collective intelligence: the role of team structure and task interdependence. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1495–1504.
[42] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293.
[43] Michel Krieger, Emily Margarete Stark, and Scott R Klemmer. 2009. Coordinating tasks on the commons: designing for personal goals, expertise and serendipity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1485–1494.
[44] Shyong Tony K Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl. 2011. WP: clubhouse?: an exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th international symposium on Wikis and open collaboration*. ACM, 1–10.
[45] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division.. In *CSCW*. 1035–1048.
[46] Pattie Maes, Ben Shneiderman, and Jim Miller. 1997. Intelligent software agents vs. user-controlled direct manipulation: a debate. In *CHI'97 Extended Abstracts on Human Factors in Computing Systems*. ACM, 105–106.
[47] Koray Mancuhan and Chris Clifton. 2014. Combating discrimination using bayesian networks. *Artificial intelligence and law* 22, 2 (2014), 211–238.
[48] Jessica K. Miller, Batya Friedman, Gavin Jancke, and Brian Gill. 2007. Value Tensions in Design: The Value Sensitive Design, Development, and Appropriation of a Corporation's Groupware System. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP '07)*. ACM, New York, NY, USA, 281–290. https://doi.org/10.1145/1316624.1316668
[49] Tim Miller. 2017. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269* (2017).
[50] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 839–848.
[51] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.
[52] David R Musicant, Yuqing Ren, James A Johnson, and John Riedl. 2011. Mentoring in Wikipedia: a clash of cultures. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM, 173–182.
[53] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.
[54] Andrew Orlowski. 2014. Jimbo tells Wikipedians: You CAN'T vote to disable 'key software features'. (2014). http://www.theregister.co.uk/2014/09/03/wales_aura_wears_off_as_jimbo_tells_wikipedians_use_what_your_told/
[55] Roy D Pea. 1986. User centered system design: New perspectives on human-computer interaction. (1986).
[56] Martin Potthast, Benno Stein, and Maik Anderka. 2008. A Wikipedia-based multilingual retrieval model. In *European conference on information retrieval*. Springer, 522–530.
[57] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.
[58] Yuqing Ren, F Maxwell Harper, Sara Drenner, Loren Terveen, Sara Kiesler, John Riedl, and Robert E Kraut. 2012. Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *Mis Quarterly* (2012), 841–864.
[59] Yuqing Ren, Robert Kraut, and Sara Kiesler. 2007. Applying common identity and bond theory to design of online communities. *Organization studies* 28, 3 (2007), 377–408.
[60] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 175–186.
[61] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[62] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.
[63] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. ACM, 285–295.
[64] Gunar Schirner, Deniz Erdogmus, Kaushik Chowdhury, and Taskin Padir. 2013. The future of human-in-the-loop cyber-physical systems. *Computer* 46, 1 (2013), 36–45.
[65] Amartya Kumar Sen. 2009. *The idea of justice*. Harvard University Press.
[66] Tom Simonite. 2013. The decline of Wikipedia. *Technology Review* 116, 6 (2013), 50–56.
[67] Koen Smets, Bart Goethals, and Brigitte Verdonk. 2008. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *AAAI workshop on Wikipedia and artificial intelligence: An Evolving Synergy*. 43–48.
[68] Bongwon Suh, Gregorio Convertino, Ed H Chi, and Peter Pirolli. 2009. The singularity is not near: slowing growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM, 8.
[69] Petr Svab. 2015. Wikipedia's Crisis of Identity. (2015). https://www.theepochtimes.com/wikipedias-crisis-of-identity_1288751.html
[70] Sebastian Thrun. 1995. Learning to play the game of chess. In *Advances in neural information processing systems*. 1069–1076.
[71] Joseph Turow. 2013. How Should We Think About Audience Power in the Digital Age? *The International Encyclopedia of Media Studies* (2013).
[72] Jennifer Valentino-Devries, Jeremy Singer-Vine, and Ashkan Soltani. 2012. Websites vary prices, deals based on users' information. *Wall Street Journal* 10 (2012), 60–68.
[73] 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*. ACM, 440.
[74] Morten Warncke-Wang, Vladislav R Ayukaev, Brent Hecht, and Loren G Terveen. 2015. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 743–756.
[75] Morten Warncke-Wang, Vivek Ranjan, Loren G Terveen, and Brent J Hecht. 2015. Misalignment Between Supply and Demand of Quality Content in Peer Production Communities.. In *ICWSM*. 493–502.
[76] Morten Warncke-Wang, Anuradha Uduwage, Zhenhua Dong, and John Riedl. 2012. In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM, 20.
[77] Bowen Yu, Yuqing Ren, Loren Terveen, and Haiyi Zhu. 2017. Predicting member productivity and withdrawal from pre-joining attachments in online production groups. In *2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017*. Association for Computing Machinery.
[78] Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 935–944.
[79] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 493–502.
[80] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089.