ORES: Facilitating re-mediation of Wikipedia's socio-technical problems

Aaron Halfaker
Wikimedia Foundation
San Francisco, CA, USA
ahalfaker@wikimedia.org

R. Stuart Geiger
University of California, Berkeley
Berkeley, CA, USA
stuart@stuartgeiger.com

Jonathan T. Morgan
Wikimedia Foundation
San Francisco, CA, USA
jmorgan@wikimedia.org

Amir Sarabadani
Wikimedia Deutschland
Berlin, Germany
amir.sarabadani@wikimedia.de

Adam Wight
Wikimedia Foundation
San Francisco, CA, USA
awight@wikimedia.org

ABSTRACT

Algorithmic systems—from rule-based bots to machine learning classifiers—have a long history of supporting the essential work of content moderation and other curation work in peer production projects. From counter-vandalism to task routing, basic machine prediction has allowed open knowledge projects like Wikipedia to scale to the largest encyclopedia in the world, while maintaining quality and consistency. However, conversations about what quality control should be and what role algorithms should play have generally been led by the expert engineers who have the skills and resources to develop and modify these complex algorithmic systems. In this paper, we describe ORES: an algorithmic scoring service that supports real-time scoring of wiki edits using multiple independent classifiers trained on different datasets. ORES decouples three activities that have typically all been performed by engineers: choosing or curating training data, building models to serve predictions, and developing interfaces or automated agents that act on those predictions. This meta-algorithmic system was designed to open up socio-technical conversations about algorithmic systems in Wikipedia to a broader set of participants. In this paper, we discuss the theoretical mechanisms of social change ORES enables and detail case studies in participatory machine learning around ORES from the 3 years since its deployment.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/0000001.0000001

CCS CONCEPTS

• **Networks** → **Online social networks**; • **Computing methodologies** → **Supervised learning by classification**; • **Applied computing** → **Sociology**; • **Software and its engineering** → **Software design techniques**; • **Computer systems organization** → Cloud computing;

KEYWORDS

Wikipedia, Reflection, Systems, Machine learning, Transparency, Fairness, Successor, Margin, Algorithms, Governance, Articulation

ACM Reference Format:
Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, Amir Sarabadani, and Adam Wight. 2019. ORES: Facilitating re-mediation of Wikipedia's socio-technical problems. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article Under review, 18 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Wikipedia—the free encyclopedia that anyone can edit—faces many challenges in maintaining the quality of its articles and sustaining the volunteer community of editors. The people behind the hundreds of different language versions of Wikipedia have long relied on automation, bots, expert systems, recommender systems, human-in-the-loop assisted tools, and machine learning to help moderate and manage content at massive scales. The issues around artificial intelligence in Wikipedia are as complex as those facing other large-scale user-generated content platforms like Facebook, Twitter, or YouTube, as well as traditional corporate and governmental organizations that must make and manage decisions at scale. And like in those organizations, Wikipedia's automated classifiers are raising new and old issues about truth, power, responsibility, openness, and representation.

Yet Wikipedia's approach to AI has long been different than in corporate or governmental contexts typically discussed in emerging fields like Fairness, Accountability, and Transparency in Machine Learning (FATML) or Critical Algorithms Studies (CAS). The volunteer community of editors has strong ideological principles of openness, decentralization, and consensus-based decision-making. The paid staff at the non-profit Wikimedia Foundation—which legally owns and operates the servers—are not tasked with making editorial decisions about content[1]. This is instead the responsibility of the volunteer community, where a self-selected set of developers build tools, bots, and advanced technologies in broad consultation with the community. Even though Wikipedia's longstanding socio-technical system of algorithmic governance is far more open, transparent, and accountable than most platforms operating at Wikipedia's scale, ORES[2][3]), the system we present in this paper, pushes even further on the crucial issue of who is able to participate in the development and use of advanced technologies.

ORES represents several innovations in openness in machine learning, particularly in seeing openness as a socio-technical challenge that is as much about scaffolding support as it is about open-sourcing code and data. With ORES, volunteers can curate labeled training data from a variety of sources for a particular purpose, commission the production of a machine classifier based on particular approaches and parameters, and make this classifier available via an API which anyone can query to score any edit to a page—operating in real time on the Wikimedia Foundation's servers. Currently, 78 classifiers have been produced for 37 languages classify edits in real-time based on criteria like "damaging / not damaging," "good faith / bad faith," or a wholistic article quality scale. ORES intentionally does not seek to produce a single classifier to enforce a gold standard of quality, nor does it prescribe particular ways in which scores and classifications will be incorporated into fully automated bots and semi-automated editing interfaces. Instead, ORES was built as a kind of cultural probe to support an open-ended set of community efforts to re-imagine what machine learning in Wikipedia is and who it is for.

Audiences for this work

The issue of open participation in machine learning raises many issues that are widely relevant to both researchers of peer production platforms like Wikipedia, as well as those working across CSCW, social computing, machine learning, and critical algorithms studies.

To researchers of CSCW systems, this paper discusses the design and role of a technical system that supports a novel type of collaborative meta-work, as ORES makes it possible for volunteers to commission the production of machine learning classifiers that other editors can use to support a variety of collaborative work practices in an established community of practice. In this paper, we detail this system's design as it was built to align with the particular ways in which volunteers work in Wikipedia. We also describe how ORES has altered the meta-work of Wikimedia tool developers.

To the FATML/CAS communities, we are introducing an open-by-design advanced algorithmic platform that is widely used to maintain a critical information resource. This platform and its context implement several of the dominant recommendations for algorithmic system builders around transparency and community consent[5, 6, 13, 32, 35]. Through the deployment of this system and subsequent design iterations, we are able to discuss novel practical considerations for what openness, accountability, and transparency mean in a large scale, real world system.

To algorithmic system-builders, we describe how we have approached key issues in developing a working, scalable, and robust system that matches the decentralized work practices of end-users in Wikipedia. Some of these approaches apply well described techniques (e.g. distributed processing and caching) while others are novel strategies for giving tool developers and their users flexibility over how to use and understand ORES's algorithmic predictions (e.g. model interrogation and threshold optimization).

In this paper, we first review related literature around open algorithmic systems, then discuss the socio-technical context of Wikipedia and the design rationale that lead us to building ORES. Next, we describe how we engineered the ORES system to match Wikipedian work practices – including innovations we've made with regards to algorithmic *openness* and *transparency*. Then we present a small set of case studies of uses and critiques of ORES' predictions that demonstrate the effectiveness of the system in meeting our design goals. Finally, we conclude with a discussion of the issues raised by this work with our target audiences: CSCW researchers, FATML/CAS researchers, social-computing researchers, and algorithmic system-builders.

[1]Except in rare cases, such as content that violates U.S. law, see http://enwp.org/WP:OFFICE
[2]https://ores.wikimedia.org
[3]http://enwp.org/:mw:ORES

2 RELATED WORK

The politics of algorithms

Algorithmic systems play increasingly cruical roles in the governance of social processes[13]. In online spaces, these systems help us deal with information overload problems: What search results best balance *relevance* and *importance*? Which books are most *related* to the ones a user likes? In other spaces, algorithmic systems help institutions run more efficiently: Who is least *risky* to loan money to? Where should police patrol to mitigate the most *dangerous* crimes? Software algorithms are increasingly used in answering such questions that have no single right answer and where prior human decisions used as training data can be problematic [2, 35].

Algorithms designed to support work change people's work practices, shifting how, where, and by whom work is accomplished[5, 13, 40]. Software algorithms gain political relevance on par with other process-mediating artifacts (e.g. laws[26]). This increasing relevance of algorithms in social and political life has renewed focus on questions of fairness and transparency[4].

There are repeated calls to address the power dynamics at play in algorithmic bias through transparency and accountability of the algorithms that govern public life and access to resources[7, 32]. The field around effective transparency and accountability mechanisms is growing. We cannot fully address the scale of concerns in this rapidly shifting literature, but we find inspiration in Kroll et al's discussion of the potential and limitations of auditing and transparency[25] and Geiger's call to go "beyond opening up the black box" [10].

This paper discusses a specific organizational and political context – Wikipedia's algorithmic quality control and socialization practices – and the development of novel algorithmic systems for support of these processes. We implement a meta-algorithmic intervention aligned with Wikipedians' principles and practices: deploying a set of prediction algorithms as a service and leaving decisions about appropriation to our users and other technology developers. Instead of seeking to train the single best classifier and implement it in our own designs, we embrace public auditing and re-interpretations of our models' predictions as an *intended* and *desired* outcome. Extensive work on technical and social ways to achieve fairness and accountability generally do not discuss this kind of infrastructural intervention on communities of practice and their articulation work.

[4]See also https://www.fatml.org/ for a conference devoted to these questions

Machine prediction in support of open production

Open peer production systems have a long history of using machine learning in service of efficiency in content moderation and task management. For Wikipedia and related Wikimedia projects, vandalism detection and quality control has been paramount for practitioners and researchers. Article quality prediction models have also been explored and applied to help Wikipedians focus their work in the most beneficial places.

Vandalism detection. The damage detection problem in Wikipedia is one of great scale. English Wikipedia receives about 160,000 new edits every day, which immediately go live without review. Wikipedians embrace this risk as the nature of an open encyclopedia, but work tirelessly to maintain quality. Every damaging or offensive edit puts the credibility of the community and their product at risk, so all edits must be reviewed as soon as possible[12].

As an information overload problem, filtering strategies using machine learning models have been developed to support the work of Wikipedia's patrollers (see [1] for an overview). In some cases, researchers directly integrated their prediction models into specific, purpose-designed tools for Wikipedians to use (e.g. STiki[39], a classifier-supported human-computation tool). Through the use of these machine learning models and boundary patrolling, most damaging edits are reverted within seconds of when they are saved[11].

Task routing. Task routing in Wikipedia is supported by a natural dynamic: people read what they are interested in, and when they see an opportunity to contribute, they do. This leads to a demand-driven contribution pattern where the most viewed content tends to be edited to the highest quality[23]. There are still many cases where Wikipedia remains misaligned[38], and content coverage biases creep in (e.g. for a long period of time, the coverage of women scientists in Wikipedia lagged far behind the rest of the encyclopedia[17]). By aligning interests with missed opportunities for contribution, these misalignments and gaps can be re-aligned and filled. Past work has explored collaborative recommender-based task routing strategies (see SuggestBot[4]), which show good success. Recently, the maintainers of SuggestBot have developed article quality prediction models to help route attention to important, but low quality articles[36]. Warncke-Wang and Halfaker have also used the article quality model to perform some one-off analyses to help Wikipedians critique and update their own manual quality assessments[37].

The Rise and Decline: Wikipedia's socio-technical problems

While Wikipedians have successfully algorithmic quality control support systems to maintain Wikipedia, a line of critical research has studied the unintended consequences of this complex socio-technical system, particularly on newcomer socialization [18, 19, 28]. In summary, Wikipedians struggled with the issues of scaling when the popularity of Wikipedia grew exponentially between 2005 and 2007[18]. In response, they developed quality control processes and technologies that prioritized efficiency by using machine prediction models[19] and templated warning messages[18]. This transformed newcomer socialization from a primarily human and welcoming activity to one that is more dismissive and impersonal[28] and cause in a steady decline in Wikipedia's editing population. The efficiency of quality control work and the elimination of damage was considered extremely politically important, while the positive experience of newcomers was less politically important.

After the research about this systemic issue came out, the political importance of newcomer experience was raised substantially. But despite targeted efforts and shifts in perception among some members of the Wikipedia community[28, 30][5], the quality control processes that were designed over a decade ago remains largely unchanged[19].

3 DESIGN RATIONALE

In this section, we discuss systemic mechanisms behind Wikipedia's socio-technical problems and how we as system builders designed ORES to have impact within Wikipedia. Past work has demonstrated how Wikipedia's problems are systemic and caused in part to inherent biases in the system of quality control in Wikipedia. To responsibly use machine learning in addressing these problems, we examined how Wikipedia functions as a distributed system using the concept of genre ecologies, focusing on how processes, policies, power, and software come together to make Wikipedia happen.

Making change in an decentralized ecology

As previously discussed, several initiatives were created to improve Wikipedian socialization practices, inlcuding the Teahouse and outreach efforts like Inspire Campaigns[27], which elicited ideas from contributors on the margins of the community. However, the process of quality control has remained largely unchanged. This assemblage of mindsets, policies, practices, and software prioritizes quality/efficiency and does so effectively [11][19] but at a cost.

Instead of pursuing the tempting technical solutions to *just fix quality control*, it is not at all apparent what better quality control would look like. Even if we did, how does one cause systemic change in a decentralized system like Wikipedia? We draw from standpoint epistemology, specifically Sandra Harding and Donna Haraway's concept of *successors*[21][22], which helps us reflect on the development of new software/process/policy components. Past work has explored developing a successor view that prioritizes the standpoints of mentors in support of new editors in Wikipedia, rather than the standpoints of vandal fighters focused on the efficiency of quality control[19][9]. However, a single point rarely changes the direction of an entire conversation or the shape of an entire ecology, so change is still elusive.

From these efforts, we know there is general interest in balancing quality/efficiency and diversity/welcomingness more effectively. So where are these designers who incorporate this expanded set of values? How do we help them bring forward their alternatives? How do we help them re-mediate Wikipedia's policies and values through their lens? How do we support the development of more successors, who can build interfaces, tools, and bots based on different ideas of what machine learning is and what it should be used for?

Our goal: Expanding the margins for successors

Successors come from the margin: they represent non-dominant values and engage in the re-mediation of articulation[29]. In our view, such successors are a primary means to change in an open genre ecology like Wikipedia. For anyone looking to enact a new view of quality control into the designs of a software system, there is a high barrier to entry: the development of a realtime machine prediction model. Without exception, all of the critical, high efficiency quality control systems that keep Wikipedia clean of vandalism and other damage employ a machine prediction model for highlighting the edits that are most likely to be bad. For example, Huggle and STiki[6] use machine prediction models to highlight likely damaging edits for human reviews. ClueBot NG[7] uses a machine prediction model to automatically revert edits that are highly likely to be damaging. These automated tools and their users work to employ a multi-stage filter that quickly and efficiently addresses vandalism[11].

[5] See also a team dedicated to supporting newcomershttp://enwp.org/:m:Growthteam

[6] http://enwp.org/WP:STiki

[7] http://enwp.org/User:ClueBot_NG

Wikipedians have long had extensive discussions and debates about the development of the thousands of relatively simple rule-based bots that are tasked with enforcing rules or supporting various tasks [8]. In contrast, there are high barriers to entry around machine learning classification models for quality control, both in knowing how they work and how to develop and operate them at Wikipedia's scale. Without these skills, it was not possible for the average Wikipedian to create an alternative view of what quality controls should be, while also accounting for efficiency and the need to scale. Notably, one of the key interventions in this area that did do so was also built by a computer scientist[19].

The result is a dominance of a certain type of individual: a computer scientist or software engineer with an eye towards improving the efficiency of quality control. This high barrier to entry and in-group effect has exacerbated the minimization of the margin and a supreme dominance of the authority of quality control regimes that were largely developed in 2006—long before the social costs of efficient quality control were understood. Worse, this barrier stands in the way of a key aspect of ecological health: diversity. We believe this lack of diversity has limited the adaptive capacity of Wikipedia's process ecology around quality management this has lead to the well-documented, long-standing issues with newcomer socialization[18].

What success looks like: Lowering barriers

Wikipedia's quality control processes are open to the development of successor systems for re-mediating quality control, but only for those with the right skills and capacities, which are not evenly distributed. We have two options for expanding the margins: (1) increase general literacy around machine classification techniques and operations at scale; or (2) minimize the need to navigate the technicalities of machine learning at scale in order to develop advanced algorithmic technologies.

Our goal in the development of ORES is to explore the second option. By deploying a high-availability machine prediction service that supports multiple classifiers at scale, designing accessible interfaces to engage with such classifiers in various ways, and engaging in basic outreach efforts, we seek to dramatically lower the barriers to the development of new algorithmic systems that could implement radically new ideas about what should be classified, how it should be classified, and how classifications and scores should be used. By enabling alternative visions of what quality control and newcomer socialization in Wikipedia should look like, we also open the doors to participation of alternative views in the genre ecology around quality control. For us, we measure success not through higher rates of precision and recall, but instead though the new conversations about how algorithmic tools affect editing dynamics, as well as new types of tools that take advantage of these resources, implementing alternative visions of what Wikipedia is and ought to be.

4 THE ORES SYSTEM

ORES has been iteratively engineered to meet the needs of Wikipedia editors and the tools that support their work. In this section, we describe how ORES' architecture was built to meet the needs of Wikipedian work processes.

Conceptual architecture

At the core, ORES is a collection of machine classifier models and an API. These models are designed and engineered by a varied set of model builders (some external researchers and others by our own engineering team) using varied sources of *training data*. The models that ORES hosts are engineered to support Wikipedian processes related to damage-detection, quality-assessment, and topic-routing, but the system is adaptable to a wide range of other models.

To make these models available for users, ORES implements a simple container service where the "container," referred to as a *ScoringModel*, represents a fully trained and tested prediction model. All *ScoringModels* contain metadata about when the model was train/tested and code for features extraction. All predictions take the form of a JSON document. The ORES service provides access to ScoringModels via a RESTful HTTP interface and serves the predictions (JSON documents) to users. We chose this service structure because Wikimedian tool developers (our target audience) are familiar with this RESTful API/JSON workflow due to the dominant use of the MediaWiki API among tool developers. See sections A and A for details and examples of about ORES' outputs.

Scaling & robustness

To be useful for Wikipedians and tool developers, ORES uses distributed computation strategies to provide a robust, fast, high-availability service. Reliability is a critical concern in Wikipedian quality control work. Interruptions in Wikipedia's algorithmic systems have historically led to increased burdens for human workers and a higher likelihood that readers will see vandalism[11]. Further, ORES needs to scale to be able to be used in multiple different tools across different language Wikipedias where its predecessors only needed to scale for use in a single tool.

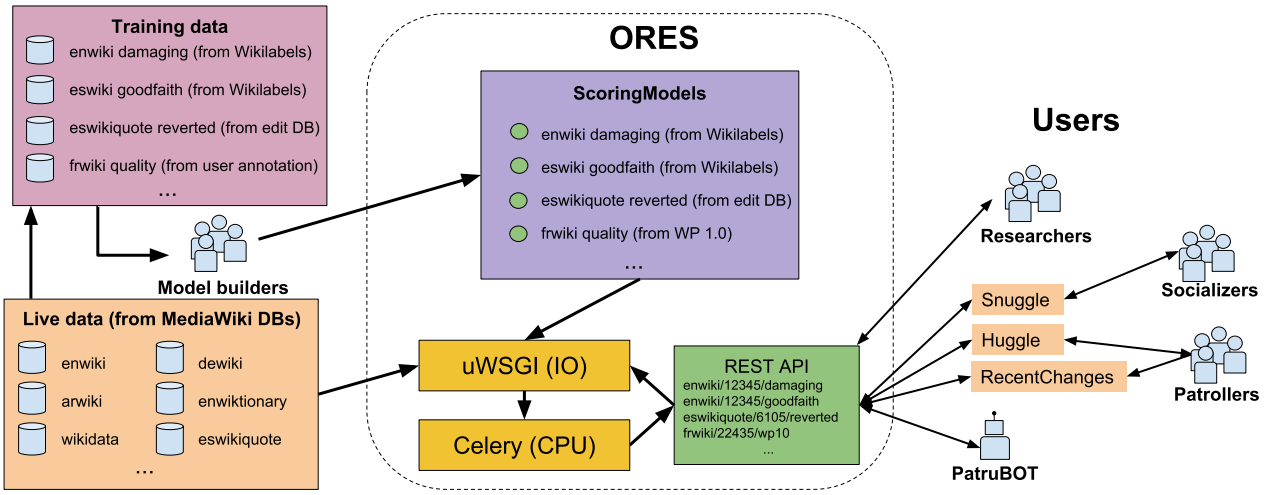


Figure 1: ORES conceptual overview. Model builders design process for training ScoringModels from training data. ORES hosts ScoringModels and makes them available to researchers and tool developers.

This horizontal scaleability is achieved in two ways: input-output (IO) workers (uwsgi[8]) and the computation (CPU) workers (celery[9]). Requests are split across available IO workers, and all necessary data is gathered using external APIs (e.g. the MediaWiki API[10]). The data is then split into a job queue managed by *celery* for the CPU-intensive work. This efficiently uses available resources and can dynamically scale, adding and removing new IO and CPU workers in multiple datacenters as needed. This is also fault-tolerant, as servers can fail without taking down the service as a whole.

Real-time processing

The most common use case of ORES is real-time processing of edits to Wikipedia immediately after they are saved. For example, those using counter-vandalism tools like Huggle monitor edits within seconds of when they are made. It is critical that ORES return these requests in a timely manner. We implement several strategies to optimize this request pattern.

Single score speed. In the worst case scenario, ORES is generating a score from scratch. This is the common case when a score is requested in real-time—which invariably occurs right after the target edit or article is saved. We work to ensure that the median score duration is around 1 second so that counter-vandalism efforts are not substantially delayed(c.f. [**?**]). Our metrics tracking currently suggests that for the week April 6-13th, 2018, our median, 75%, and 95% score response timings are 1.1, 1.2, and 1.9 seconds respectively.

Caching and precaching. In order to take advantage of our users' overlapping interests in scoring recent activity, we also maintain a basic least-recently-used (LRU) cache[11] using a deterministic score naming scheme (e.g. `enwiki:123456:damaging` would represent a score needed for the English Wikipedia damaging model for the edit identified by 123456). This allows requests for scores that have recently been generated to be returned within about 50ms via HTTPS. In other words, a request for a recent edit that had previously been scored is 20X faster due to this cache.

In order to make sure that scores for *all recent edits* are available in the cache for real-time use cases, we implement a "precaching" strategy that listens to a high-speed stream of recent activity in Wikipedia and automatically requests scores for a specific subset of actions (e.g. edits). With our LRU and precaching strategy, we consistently attain a cache hit rate of about 80%.

De-duplication. In real-time ORES use cases, it's common to receive many requests to score the same edit/article right after it was saved. We use the same deterministic score naming scheme from the cache to identify scoring tasks, and ensure that simultaneous requests for that same score are de-duplicated. This allows our service to trivially scale to support many different robots and tools on the same wiki.

[8] https://uwsgi-docs.readthedocs.io/
[9] http://www.celeryproject.org/
[10] http://enwp.org/:mw:MW:API
[11] Implemented natively by Redis, https://redis.io

Batch processing

Many different types of Wikipedia's bots rely on periodic, batch processing strategies to support Wikipedian work processes[8]. For example, many bots are designed to build worklists for Wikipedia editors (e.g. [4]) on a daily or weekly basis, and many of these tools have adopted ORES to include an article quality prediction for use in prioritization of work (see section 6). Work lists are either built from the sum total of all 5m+ articles in Wikipedia, or from some large subset specific to a single WikiProject (e.g. WikiProject Women Scientists claims about 6k articles[12]). We've observed robots submitting large batch processing jobs to ORES once per day. It's relevant to note that many researchers are also making use of ORES for various analyses, and their activity usually shows up in our logs as a similar burst of requests.

In order to most efficiently support this type of querying activity, we implemented batch optimizations in ORES by splitting IO and CPU operations into distinct stages. During the IO stage, all data is gathered for all relevant scoring jobs in batch queries. During the CPU stage, scoring jobs are split across our distributed processing system. This batch processing affords up to a 5X increase in time to scoring speed for large requests[33]. At this rate, a user can request 10s of million of scores in less than 24 hours in the worst case scenario (no scores were cached) without substantially affecting the service for others.

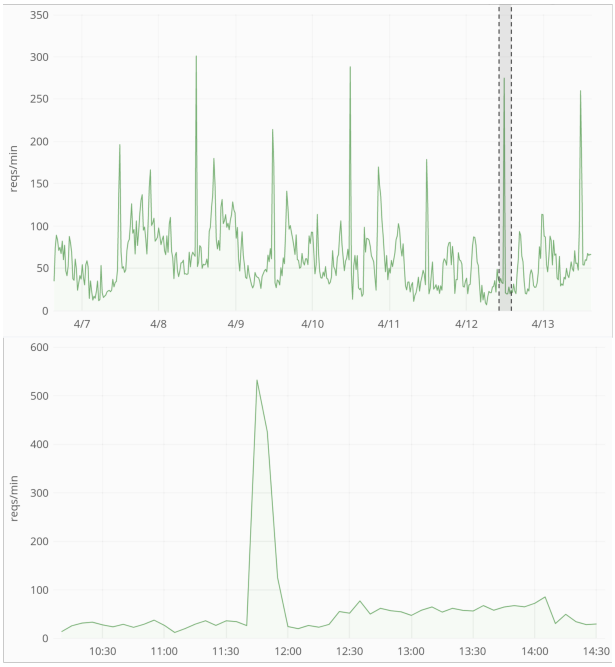
Empirical access patterns

The ORES service has been online since July 2015[20]. Since then, usage has steadily risen as we've developed and deployed new models and additional integrations are made by tool developers and researchers. Currently, ORES supports 78 different models and 37 different language-specific wikis.

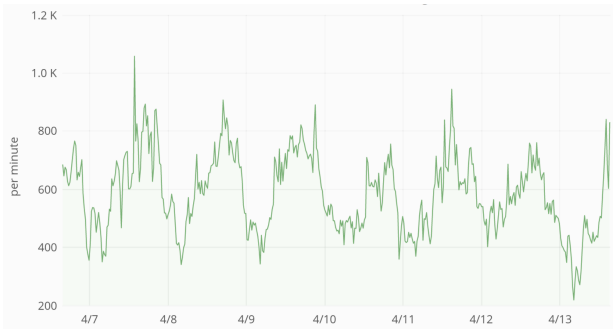
Generally, we see 50 to 125 requests per minute from external tools that are using ORES' predictions (excluding the MediaWiki extension that is more difficult to track). Sometimes these external requests will burst up to 400-500 requests per second. Figure 2a shows the periodic and "bursty" nature of scoring requests received by the ORES service. For example, every day at about 11:40 UTC, the request rate jumps—most likely a batch scoring job such as a bot.

Figure 2b shows the rate of precaching requests coming from our own systems. This graph roughly reflects the rate of edits that are happening to all of the wikis that we support since we'll start a scoring job for nearly every edit as it happens. Note that the number of precaching

[12]As demonstrated by https://quarry.wmflabs.org/query/14033



(a) External requests per minute with a 4 hour block broken out to highlight a sudden burst of requests



(b) Precaching requests per minute

Figure 2: Request rates to the ORES service for the week ending on April 13th, 2018

requests is about an order of magnitude higher than our known external score request rate. This is expected, since Wikipedia editors and the tools they use will not request a score for every single revision. This is a computational price we pay to attain a high cache hit rate and to ensure that our users get the quickest possible response for the scores that they *do* need.

Taken together these strategies allow us to optimize the real-time quality control workflows and batch processing jobs of Wikipedians and their tools. Without serious effort to make sure that ORES is practically fast and highly available to real-time use cases, ORES would become irrelevant to the target audience and thus

irrelevant as a boundary-lowering intervention. By engineering a system that conforms to the work-process needs of Wikipedians and their tools, we've built a systems intervention that has the potential gain wide adoption in Wikipedia's technical ecology.

5 INNOVATIONS IN OPENNESS

We developed ORES in the context of Wikipedia, an egalitarian, decentralized, and radically transparent community. So with ORES, we sought to maintain these values in our system design and model building strategies. The flow of data, from random samples through model training, evaluation, and application, is open for review, critique, and iteration. Further, we have developed novel strategies for opening ORES models to play and experimentation based on user requests. In this section, we describe some of the key, novel innovations that have made ORES fit Wikipedian concerns and be flexible to re-appropriation. The appendix also contains information about ORES' detailed prediction output (section A), how users and tools can adjust their use to model fitness (sections A and A), and how the whole model development workflow is made inspectable and replicable (section A).

Collaboratively labeled data

There are two primary strategies for gathering labeled data for ORES' models: found traces and manual labels.

Found traces. For many models, the MediaWiki platform records a rich set of digital traces that can be assumed to reflect a useful human judgement. For example, in Wikipedia, it is very common that damaging edits will eventually be reverted and that good edits will not be reverted. Thus the revert action (and remaining traces) can be used to assume that the reverted edit is damaging. We have developed a re-usable script[13] that when given a sample of edits, will label the edits as "reverted_for_damage" or not based on a set of constraints: edit was reverted within 48 hours, the reverting editor was not the same person, and the edit was not restored by another editor.

However, this "reverted_for_damage" label is problematic in that many edits are reverted not because they are damaging but because they are involved in some content dispute. Operationalizing quality by exclusively measuring what persists in Wikipedia reinforces Wikipedia's well-known systemic biases, which is a similar problem in using found crime data in predictive policing. Also, the label does not differentiate damage that is a good-faith mistake from damage that is intentional vandalism. So in the case of damage prediction models, we only make use of the "reverted_for_damage" label when manually labeled data is not available.

Manual labeling campaigns with Wiki Labels. We hold manual labeling by human Wikipedians as the gold standard for purposes of training a model to replicate human judgement. By asking Wikipedians to demonstrate their judgement on examples from their own wikis, we can most closely tailor model predictions to match the judgements that make sense to these communities. This contrasts with found data, which deceptively appears to be a better option because of its apparent completeness: every edit was either reverted or not. Yet as previously discussed, there are many issues with bias, and the implicit signals may not exactly match the intended use of the model. Manual labeling has a high up-front expense of human labor. In order to efficiently utilize valuable time investments by our collaborators – mostly volunteer Wikipedians – we developed a system called "Wiki Labels[14]". Wiki Labels allows Wikipedians to submit judgments of specific random samples of Wiki content using a convenient interface and logging in via their Wikipedia account.

For example, to supplement our models of edit quality, we replace the models based on found "reverted_for_damage" traces with manual judgments where we specifically ask labelers to distinguish "damaging"/good from "good-faith"/vandalism. "Good faith" is a well-established term in Wikipedian culture, with specific local meanings that are different than their broader colloquial use — similar to how Wikipedians define "consensus" or "neutrality". Using these labels we can build two separate models which allow users to filter for edits that are likely to be good-faith mistakes[16], to just focus on vandalism, or to apply themselves broadly to all damaging edits.

Dependency injection and interrogability

One of the key features of ORES that allows scores to be generated in an efficient and flexible way is a dependency injection framework. We use a dependency solver to determine what data is necessary for a scoring job and eventually compute the features used by a prediction model.

The flexibility provided by the dependency injection framework lets us implement a novel strategy for exploring *how* ORES' models make predictions. By exposing the features extracted to ORES users and allowing them to inject their own features, we can allow users to ask how predictions would change if the world were different. Let's say you wanted to explore how ORES judges

[13] see *autolabel* in https://github.com/wiki-ai/editquality

[14] http://enwp.org/:m:Wikilabels

unregistered (anon) editors differently from registered editors. Figure 3 demonstrates two prediction requests to ORES.

Figure 3a shows that ORES' "damaging" model concludes that the edit identified by the *revision ID* of 34234210 is not damaging with 93.9% confidence. We can ask ORES to make a prediction about the exact same edit, but to assume that the editor was unregistered (anon). Figure 3b shows the prediction if edit were saved by an anonymous editor. ORES would still conclude that the edit was not damaging, but with less confidence (91.2%). By following a pattern like this for a single edit or a set of edits, we can get to know how ORES prediction models account for anonymity through experience with practical examples.

Interrogability has also been used in creative new uses beyond bias explorations. Some of our users have levered the feature injection system to expose *hypothetical* predictions to support their work. See the discussion of Ross's work recommendation tools in Section 6.

6 ADOPTION PATTERNS

When we designed and developed ORES, we were targeting a specific problem: expanding the set values applied to the design of quality control tools to include recent a recent understanding of the importance of newcomer socialization. We do not have any direct control of how developers chose to use ORES. We hypothesize that, by making edit quality predictions available to all developers, we would lower the barrier to experimentation in this space. From our experiences, it is clear that we lowered barriers to experimentation. After we deployed ORES, we implemented some basic tools to showcase ORES, but we observed a steady adoption of our various prediction models by external developers in current tools and through the development of new tools.[15]

Adoption in current tools

Many tools for counter-vandalism in Wikipedia were already available when we developed ORES. Some of them made use of machine prediction (e.g. Huggle[16], STiki, ClueBot NG), but most did not. Soon after we deployed ORES, many developers that had not previously included their own prediction models in their tools were quick to adopt ORES. For example, RealTime Recent Changes[17] includes ORES predictions along-side their realtime interface and FastButtons,[18] a Portuguese Wikipedia gadget, began displaying ORES predictions next to their buttons for quick reviewing and reverting damaging edits.

Other tools that were not targeted at counter-vandalism also found ORES predictions—specificly that of *article quality* (wp10)—useful. For example, RATER,[19] a gadget for supporting the assessment of article quality began to include ORES predictions to help their users assess the quality of articles and SuggestBot,[20][4] a robot for suggesting articles to an editor, began including ORES predictions in their tables of recommendations.

Development of new tools

Many new tools have been developed since ORES was released that may not have been developed at all otherwise. For example, the Wikimedia Foundation product department developed a complete redesign on MediaWiki's Special:RecentChanges interface that implements a set of powerful filters and highlighting. They took the ORES Review Tool to it's logical conclusion with an initiative that they referred to as Edit Review Filters.[21] In this interface, ORES scores are prominently featured at the top of the list of available features, and they have been highlighted as one of the main benefits of the new interface to the editing community.

When we first developed ORES, English Wikipedia was the only wiki that we are aware of that had a robot that used machine prediction to automatically revert obvious vandalism[3]. After we deployed ORES, several wikis developed bots of their own to use ORES predictions to automatically revert vandalism. For example, PatruBOT in Spanish Wikipedia[22] and Dexbot in Persian Wikipedia[23] now automatically revert edits that ORES predicts are damaging with high confidence. These bots have been received with mixed acceptance. Because of the lack of human oversight, concerns were raised about PatruBOT's false positive rate but after consulting with the developer, we were able to help them find an acceptable threshold of confidence for auto-reverts.

One of the most noteworthy new applications of ORES is the suite of tools developed by Sage Ross to support the Wiki Education Foundation's[24] activities. Their organization supports classroom activities that involve editing Wikipedia. They develop tools and dashboards that help students contribute successfully and to help teachers

[15] See complete list: http://enwp.org/:mw:ORES/Applications
[16] Notably, Huggle adopted ORES prediction models soon after we deployed
[17] http://enwp.org/:m:RTRC
[18] http://enwp.org/:pt:Wikipdia:Scripts/FastButtons
[19] http://enwp.org/::en:WP:RATER
[20] http://enwp.org/User:SuggestBot
[21] http://enwp.org/:mw:Edit_Review_Improvements
[22] https://es.wikipedia.org/wiki/Usuario:PatruBOT
[23] https://fa.wikipedia.org/wiki/User:Dexbot
[24] https://wikiedu.org/

```
"damaging": {
  "score": {
    "prediction": false,
    "probability": {
      "false": 0.938910157824447,
      "true": 0.06108984217555305
    }
  }
}
```

(a) Prediction with `anon = false` injected

```
"damaging": {
  "score": {
    "prediction": false,
    "probability": {
      "false": 0.9124151990561908,
      "true": 0.0875848009438092
    }
  }
}
```

(b) Prediction with `anon = true` injected

Figure 3: Two "damaging" predictions about the same edit are listed for ORES. In one case, ORES is asked to make a prediction assuming the editor is unregistered (anon) and in the other, ORES is asked to assume the editor is registered.

monitor their students' work. Ross has recently published about how he interprets meaning from ORES' article quality models[31] (an example of re-appropriation) and he has uses the article quality model in their new editor support dashboard[25] in a novel way to support new editors. Specifically, Ross's tool[26] uses our feature injection system (see Section 5) suggesting work to new editors. This system asks ORES to score a student's draft and then asking ORES to reconsider the predicted quality level of the article with *one more header*, *one more image*, or *one more citation*. In doing so, Ross built an intelligent user interface that can expose the internal structure of a model in order to recommend the most productive development to the article—the change that will most likely bring it to a higher quality level.

7 CASE STUDIES IN REFLECTION

When we first deployed ORES, we reached out to several different wiki communities and invited them to test the system for use in patrolling for vandalism. In these announcements, we encouraged editors to install ScoredRevisions, the only tool that used ORES's edit quality models at the time. ScoredRevisions both highlights edits that are likely to be damaging (as predicted by the model) and displays the likelihood of the prediction as a percentage.

Before long, our users began filing false-positive reports on wiki pages of their own design—some after our request, but mostly on their own. In this section, we describe three cases where our users independently developed these false-positive reporting pages and how they used them to understand ORES, the roles of automated quality control in their own spaces, and to communicate with us about model bias.

Report mistakes (Wikidata)

Improvements [edit]

Diff id	Damaging	Old score	Score1	Score2	Score3	1st improv.	2nd	3rd	Overall
210649590	No	91%	30%	5%	0%	+61%	+25%	+5%	+91%
237999679	No	84%	71%	63%	60%	+13%	+8%	+3%	+24%
243937491	No	95%	46%	74%	71%	+49%	-28%	+3%	+24%
251530750	No	91%	55%	56%	55%	+36%	-1%	+1%	+36%
253584599	No	99%	89%	78%	70%	+10%	+11%	+8%	+29%
257856652	No	91%	30%	4%	1%	+61%	+26%	+3%	+90%

Figure 4: A slice of the ORES report mistakes table in Wikidata.

When we first deployed prediction models for Wikidata—a free and open knowledge base that can be read and edited by both humans and machines[27]—we were breaking new ground by building a damage detection classifier based on a structured data wiki[33]. We created a page called "Report mistakes" and invited users to tell us about mistakes that the prediction model made on that page. We left the format and structure largely up to the users.

Within 20 minutes, we received our first report. As reports streamed in, we began to respond to them and make adjustments to the model building process to address data extraction bugs and to increase the signal so that the model differentiate damage from non-damaging edits. After a month of reports and bug fixes, we decided to build a table to represent the progress that we made in iterations on the model against the reported false-positives (Figure 4). Each row represents false-positive, and each column describes the progress we made in not

[25] https://dashboard-testing.wikiedu.org
[26] https://dashboard-testing.wikiedu.org
[27] https://wikidata.org

detecting those edits as damaging in subsequent iterations of the model. Through this process, we learned how Wikidata editors understood and saw damage, as well as how our modeling and feature extraction process captured signals in ways that differed from Wikidata editors' understandings. Because of this back-and-forth collaboration made possible through ORES's various features, we were able to publicly demonstrate improvements to this community.

Patrolling/ORES (Italian Wikipedia)

Italian Wikipedia was one of the first wikis where we deployed basic edit quality models. Our local collaborator, who helped us develop the language specific features, User:Rotpunkt, created a page for ORES[28] with a section for reporting false-positives ("falsi positivi"). Within several hours, Rotpunkt and a few other editors noticed some trends in their false positive reports. These editors began to collect false positives under different headers representing themes they were seeing. Through this process, editors from Italian Wikipedia were effectively performing an inductive, grounded theory-esque exploration ORES errors, trying to identify themes and patterns in the errors that ORES was making.

One of the themes they identified fell under the header: "corrections to the verb for *have*" ("correzioni verbo avere"). It turns out that the word "ha" in Italian translates to the English verb "to have". While in Engligh and many other langauges, "ha" is laughing and adding "ha" repeatedly is a common type of vandalism seen in all langauges of Wikipedia. We'd built a common feature in the damage model called "informal words" that captured these types of patterns. But in this case, it was clear that in Italian "ha" should not carry signal while "hahaha" still should.

Because of the work of Rotpunkt and his collaborators in Italian Wikipedia, we were able to recognize the source of this issue (a set of features intended to detect the use of *informal language* in articles) and to remove "ha" from that list for Italian Wikipedia.

PatruBOT (Spanish Wikipedia)

Soon after we released support for Spanish Wikipedia, a volunteer developer made a bot to automatically revert damaging edits using ORES's predictions for the "damaging" model (PatruBOT). This bot was not running for long before our discussion pages were bombarded with confused Spanish-speaking editors asking us questions about why ORES did not like their work. We struggled to understand the origin of the complaints until someone reached out about PatruBOT and its activities.

When we examined the case, we found it was one of tradeoffs between precision/recall and false positives/negatives—a common issue with machine learning applications. We concluded that PatruBOT's threshold for reverting was too sensitive. ORES reports a classification and a probability score, but it is up to the developers to decide if, for example, the bot will only auto-revert edits classified as damage with a .90, .95, .99, or higher likelihood estimate. A higher threshold will minimize the chance a good edit will be mistakenly auto-reverted, but also increase the chance that a bad edit will not be auto-reverted. Ultimately, deciding where to draw the line between false positives and false negatives is a decision for that volunteer editing community.

The Spanish Wikipedians who were concerned with these issues began a discussion about PatruBOT's activities and blocked the bot until the issue was sorted. Using wiki pages, they organized an crowdsourced evaluation of the fitness of PatruBOT's behavior[29]. This evaluation and discussion is ongoing, [30] but it shows how stakeholders do not need to have an advanced understanding in machine learning evaluation to meaningfully participate in a sophisticated discussion about how, when, why, and under what conditions such classifiers should be used.

Bias against anonymous editors

Shortly after we deployed ORES, we received reports that ORES's damage detection models were overly biased against anonymous editors. At the time, we were using Linear SVM[31] estimators to build classifiers, and we were considering making the transition towards ensemble strategies like GradientBoosting and RandomForest estimators.[32] We took the opportunity to look for bias in the error of estimation between anonymous editors and newly registered editors. By using our feature injection/interrogation strategy (described in Section 5), we could ask our current prediction models how they would change their predictions if the exact same edit were made by a different editor.

Figure 5 shows the probability density of the likelihood of "damaging" given three different passes over the exact same test set, using two of our modeling strategies.

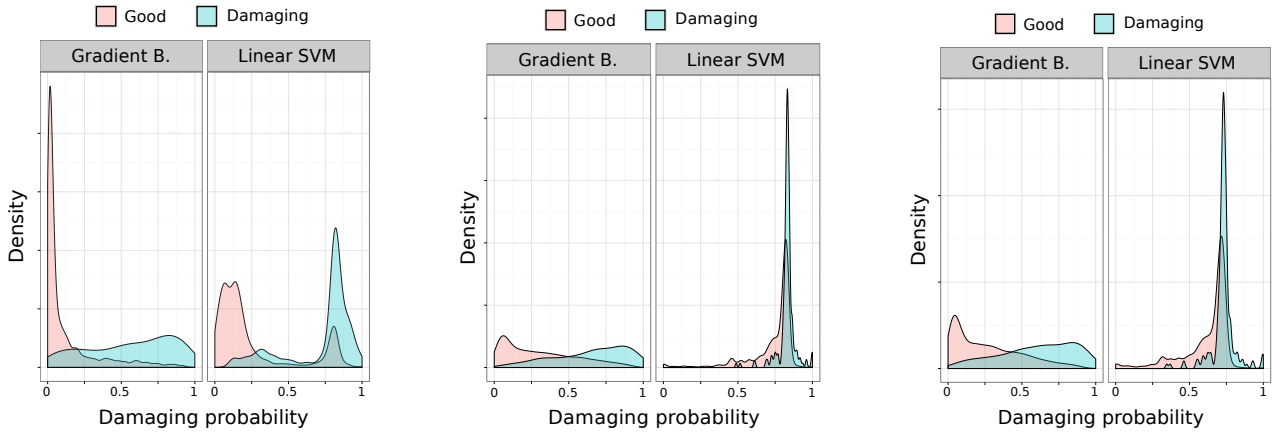
[28] https://it.wikipedia.org/wiki/Progetto:Patrolling/ORES

[29] https://es.wikipedia.org/wiki/Wikipedia:Mantenimiento/Revisi%C3%B3n_de_errores_de_PatruBOT%2FAn%C3%A1lisis

[30] https://es.wikipedia.org/wiki/Wikipedia:Caf%C3%A9%2FArchivo%2FMiscel%C3%A1nea%2FActual#Parada_de_PatruBOT

[31] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[32] http://scikit-learn.org/stable/modules/ensemble.html



(a) No injected features (b) Everyone is anonymous (c) Everyone is newly registered

Figure 5: The distributions of the probability of a single edit being scored as "damaging" based on injected features for the target user-class is presented. Note that when injecting user-class features (anon, newcomer), all other features are held constant.

Figure 5a shows that, when we leave the features to their natural values, it appears that both models are able to differentiate effectively between damaging edits (high-damaging probability) and non-damaging edits (low-damaging probability) with the odd exception of a large amount of non-damaging edits with a relatively high-damaging probability around 0.8 in the case of the Linear SVM model. Figures 5b and 5c show a stark difference. For the scores that go into these plots, characteristics of anonymous editors and newly registered editors were injected for all of the test edits. We can see that the GradientBoosting model can still differentiate damage from non-damage while the Linear SVM model flags nearly all edits as damage in both case.

Through the reporting of this issue and our subsequent analysis, we were able to identify the issue and show that an improvement to our modeling strategy mitigates the problem. Without such a tight feedback loop, we most likely would not have noticed how poorly ORES's damage detection models were performing in practice. Worse, it might have caused vandal fighters to be increasingly (and inappropriately) skeptical of contributions by anonymous editors and newly registered editors—two groups of contributors that are already met with unnecessary hostility[33] [18].

8 CONCLUSION AND FUTURE WORK

ORES as a socio-technical system has helped us 1) refine our understandings of volunteers' needs across wiki communities, 2) identify and address biases in ORES's models, and 3) reflect on how people think about what types of automation they find acceptable in their *spaces*. Through our participatory design process with various Wikipedian communities, we've arrived at several innovations in open machine learning practice that represent advancements in the field.

As we stated in Section 3, we measure success in "new conversations about how algorithmic tools affect editing dynamics, as well as new types of tools that take advantage of these resources, implementing alternative visions of what Wikipedia is and ought to be." We have demonstrated through discussion adoption patterns and case studies in reflection around the use of algorithmic systems that something fundamental is *working*. ORES is being heavily adopted. The meaning of ORES models is being re-appropriated. Both the models and the technologies that use the models are being collaboratively audited by their users and those who are affected.

Participatory machine learning

In a world dominated by for-profit social computing and user-generated content platforms—often marketed by their corporate owners as "communities" [14]—Wikipedia is an anomaly. While the non-profit Wikimedia Foundation has only a fraction of the resources as Facebook or Google, the unique principles and practices in the broad Wikipedia/Wikimedia movement are a generative constraint. ORES emerged out of this context, operating at the intersection of a pressing need to deploy efficient machine learning at scale for content moderation, but to do so in ways that enable volunteers to develop and deploy advanced technologies on their own terms. Our

[33] http://enwp.org/:en:Wikipedia:IPs_are_human_too

approach is in stark contrast to the norm in machine learning research and practice, which involves a more top-down mode of developing the most precise classifiers for a known ground truth, then deploying a complete technology for end-users, who must treat them as black boxes.

The more wiki-inspired approach to what we call "participatory machine learning" ORES supports imagines classifiers to be just as provisional and open to skeptical reinterpretation as the content of Wikipedia's encyclopedia articles. And like Wikipedia articles, we suspect some classifiers will be far better than others based on how volunteers develop and curate them, for various definitions of "better" that are already being actively debated. Our case studies briefly indicate how volunteers have collectively engaged in sophisticated discussions about how they ought to use machine learning. ORES' fully open, reproducible, and auditable code and data pipeline—from training data to models to scored predictions—enables a wide and rich range of new collaborative practices. We see ORES as a more socio-technical and specifically CSCW approach to issues around fairness, accountability, and transparency in machine learning, where much attention is placed on technical solutions, like interactive visualizations for model interpretability or mathematical guarentees of different operationalized definitions of fairness. Our approach is specific to the particular genre ecology, work practices, and values of Wikipedia, and we have shown how ORES has been developed to fit into this complex socio-technical system.

Critical reflection

In section 7, we show evidence of critical reflection on the current processes and the role of algorithms in quality control. We believe that the case studies that we describe both show that collaborative auditing is taking place and that the wide proliferation of tools that provide surprising alternative uses of ORES suggest that Wikipedians feel a renewed power over their quality control processes. We are inspired by much of the concern that has surfaced for looking into biases in ORES' prediction models (e.g. anon bias and the Italian "ha") and over what role algorithmis should have in directly reverting human actions (e.g. PatruBOT and DexBot).

Eliciting this type of critical reflection and empowering users to engage in their own choices about the roles of algorithmic systems in their social spaces has typically been more of a focus from the Critical Algorithms Studies literature (e.g. [2, 24]. This literature also emphasizes a need to see algorithmic systems as dynamic and constantly under revision by developers [34]—work that is invisible in most platforms, but foregrounded in ORES. In these case studies, we see that given ORES' open API and Wikipedia's collaborative wiki pages, Wikipedians will audit ORES' predictions and collaborate with each other to build information about trends in ORES' mistakes and how they expected their own processes to function.

Future work

Observing ORES in practice suggests avenues of future work toward crowd-based auditing tools. As our case studies suggest, auditing of ORES' predictions and mistakes has become a very popular activity. Even though we did not design interfaces for discussion and auditing, some Wikipedians have used unintended affordances of wiki pages and MediaWiki's template system to organize similar processes for flagging false positives and calling them to our attention. This process has proved invaluable for improving model fitness and addressing critical issues of bias against disempowered contributors. To better facilitate this process, future system builders should implement structured means to refute, support, discuss, and critique the predictions of machine models. With a structured way to report what machine prediction gets right and wrong, we can make it easier for tools that use ORES to also allow for reporting mistakes and for others to infer trends. For example, a database of ORES mistakes could be queried in order to build the kind of thematic analyses that Italian Wikipedians showed us. By supporting such an activity, we are working to transfer more power from ourselves and to our users. Should one of our models develop a nasty bias, our users will be more empowered to coordinate with each other, show that the bias exists and where it causes problems, and either get the model's predictions turned off or even shut down ORES (e.g. PatruBOT).

We also look forward to what future work in the space of critical algorithm studies will do with ORES. As of writing, most of the studies and critiques of *subjective algorithms*[35] focus on large for-profit organizations like Google and Facebook—organizations that can't afford to open up their proprietary algorithms due to competition. Wikipedia is one of the largest and most important information resources in the world. The algorithms that ORES makes available are part of the decision process that leads to some people's contributions remaining and others being removed. This is a context where *algorithms matter to humanity*, and we are openly experimenting with the kind of transparent and open processes that *fairness and transparency in machine learning* researchers are advocating. Yet, we have new problems and new opportunities. There is a large body of work exploring how biases manifest and how unfairness can play out in

algorithmically mediated social contexts. ORES would be an excellent place to expand the literature within a real and important field site.

Finally, we also see potential in allowing Wikipedians, the denizens of Wikipedia, to freely train, test, and use their own prediction models without our engineering team involved in the process. Currently, ORES is only suited to deploy models that are trained and tested by someone with a strong modeling and programming background. That doesn't need to be the case. We have been experimenting with demonstrating ORES model building processes using Jupyter Notebooks[34][35] and have found that beginning programmers can understand the work involved. This is still not the holy grail of crowd-developed machine prediction—where all of the incidental complexities involved in programming are removed from the process of model development and evaluation. Future work exploring strategies for allowing end-users to build models that are deployed by ORES would surface the relevant HCI issues involved and the changes the the technological conversations that such a margin-opening intervention might provide.

9 ACKNOWLEDGEMENTS

REDACTED FOR REVIEW

REFERENCES

[1] B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 277–288.

[2] Solon Barocas, Sophie Hood, and Malte Ziewitz. 2013. Governing algorithms: A provocation piece. *SSRN. Paper presented at Governing Algorithms conference.* (2013). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2245322

[3] Jacobi Carter. 2008. ClueBot and vandalism on Wikipedia. https://web.archive.org/web/20120305082714/http://www.acm.uiuc.edu/~carter11/ClueBot.pdf

[4] 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.

[5] Kate Crawford. 2016. Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values* 41, 1 (2016), 77–92.

[6] Nicholas Diakopoulos. 2015. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3, 3 (2015), 398–415.

[7] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic Transparency in the News Media. *Digital Journalism* 5, 7 (2017), 809–828. https://doi.org/10.1080/21670811.2016.1208053

[8] R Stuart Geiger. 2011. The lives of bots. In *Critical Point of View: A Wikipedia Reader*. Institute of Network Cultures, Amsterdam, 78–93. http://stuartgeiger.com/lives-of-bots-wikipedia-cpov.pdf

[9] R Stuart Geiger. 2014. Successor Systems: The Role of Reflexive Algorithms in Enacting Ideological Critique. *AoIR Selected Papers of Internet Research* 4 (2014). https://spir.aoir.org/index.php/spir/article/download/942/611

[10] R. Stuart Geiger. 2017. Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data & Society* 4, 2 (2017), 2053951717730735. https://doi.org/10.1177/2053951717730735

[11] 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.

[12] R Stuart Geiger and David Ribes. 2010. The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 117–126.

[13] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167 (2014).

[14] Tarleton Gillespie. 2018. *Custodians of the internet : platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, New Haven.

[15] Aaron Halfaker. 2016. Notes on writing a Vandalism Detection paper. http://socio-technologist.blogspot.com/2016/01/notes-on-writing-wikipedia-vandalism.html

[16] Aaron Halfaker. 2017. Automated classification of edit quality (worklog, 2017-05-04). https://meta.wikimedia.org/wiki/Research_talk:Automated_classification_of_edit_quality/Work_log/2017-05-04

[17] Aaron Halfaker. 2017. Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect. In *Proceedings of the 13th International Symposium on Open Collaboration*. ACM, 19.

[18] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedias reaction to popularity is causing its decline. *American Behavioral Scientist* 57, 5 (2013), 664–688.

[19] 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.

[20] Aaron Halfaker and Dario Taraborelli. 2015. Artificial Intelligence Service ORES Gives Wikipedians X-Ray Specs to See Through Bad Edits. https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/

[21] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.

[22] Sandra G Harding. 1987. *Feminism and methodology: Social science issues*. Indiana University Press.

[23] Benjamin Mako Hill and Aaron Shaw. 2014. Consider the redirect: A missing dimension of Wikipedia research. In *Proceedings of The International Symposium on Open Collaboration*. ACM, 28.

[34] http://jupyter.org

[35] e.g. https://github.com/wiki-ai/editquality/blob/master/ipython/reverted_detection_demo.ipynb

[24] Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (2017), 14–29. https://doi.org/10.1080/1369118X.2016.1154087
[25] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.
[26] Lawrence Lessig. 1999. *Code: And other laws of cyberspace*. Basic Books.
[27] Jonathan T. Morgan. 2015. What we learned from the Inspire campaign to increase gender diversity on Wikimedia. ttps://blog.wikimedia.org/2015/05/28/what-we-learned-from-the-inspire-campaign/
[28] 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.
[29] Gabriel Mugar. 2017. Preserving the Margins: Supporting Creativity and Resistance on Digital Participatory Platforms. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 83.
[30] Sneha Narayan, Jake Orlowitz, Jonathan T Morgan, and Aaron Shaw. 2015. Effects of a Wikipedia Orientation Game on New User Edits. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*. ACM, 263–266.
[31] Sage Ross. 2016. Visualizing article history with Structural Completeness. https://wikiedu.org/blog/2016/09/16/visualizing-article-history-with-structural-completeness/
[32] 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.
[33] Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. 2017. Building automated vandalism detection tools for Wikidata. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1647–1654.
[34] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4, 2 (2017). https://doi.org/10.1177/2053951717738104
[35] Zeynep Tufekci. 2015. Algorithms in our midst: Information, power and choice when software is everywhere. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1918–1918.
[36] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell me more: an actionable quality model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*. ACM, 8.
[37] Morten Warncke-Wang and Aaron Halfaker. 2017. Screening WikiProject Medicine articles for quality. https://meta.wikimedia.org/wiki/Research:Screening_WikiProject_Medicine_articles_for_quality
[38] 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.
[39] Andrew G West, Sampath Kannan, and Insup Lee. 2010. STiki: an anti-vandalism tool for Wikipedia using spatio-temporal analysis of revision metadata. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*. ACM, 32.
[40] Shoshana Zuboff. 1988. *In the age of the smart machine: The future of work and power*. Vol. 186. Basic books New York.

A APPENDIX

ORES system details

In this section, we describe some of the details of the ORES system.

Score documents. The predictions made by through ORES are human- and machine-readable. In general, our classifiers will report a specific prediction along with a set of probability (likelihood) for each class. By providing detailed information about a prediction, we allow users to re-purpose the prediction for their on use. Consider article quality (wp10) prediction output in Figure 7.


```
"wp10": {
  "score": {
    "prediction": "Start",
    "probability": {
      "FA": 0.00329313015, "GA": 0.0058529554,
      "B": 0.06062338048, "C": 0.01991363271,
      "Start": 0.754330134, "Stub": 0.1559867667
    }
  }
}
```


Figure 6: Result of https://ores.wikimedia.org/v3/scores/enwiki/34234210/wp10

A developer making use of a prediction like this may choose to present the raw prediction "Start" (one of the lower quality classes) to users or to implement some visualization of the probability distribution across predicted classed (75% Start, 16% Stub, etc.). They might even choose to build an aggregate metric that weights the quality classes by their prediction weight (e.g. Ross's student support interface[31] or the *weighted sum* metric from [17]).

Model information. In order to use a model effectively in practice, a user needs to know what to expect from model performance. E.g. how often is it that when an edit is predicted to be "damaging" it actually is? (*precision*) or what proportion of damaging edits should I expect will be caught by the model? (*recall*) The target metric of an operational concern depends strongly on the intended use of the model. Given that our goal with ORES is to allow people to experiment with the use and reflection of prediction models in novel ways, we sought to build an general model information strategy.

```
"damaging": {
  "type": "GradientBoosting",
  "version": "0.4.0",
  "environment": {"machine": "x86_64", ...},
  "params": {"center": true, "init": null,
             "label_weights": {"true": 10},
             "labels": [true, false],
             "learning_rate": 0.01,
             "min_samples_leaf": 1,
             ...},
  "statistics": {
    "counts": {
      "labels": {"false": 18702, "true": 743},
      "n": 19445,
      "predictions": {
        "false": {"false": 17989, "true": 713},
        "true": {"false": 331, "true": 412}}},
    "precision": {
      "labels": {"false": 0.984, "true": 0.34},
      "macro": 0.662, "micro": 0.962},
    "recall": {
      "labels": {"false": 0.962, "true": 0.555},
      "macro": 0.758, "micro": 0.948},
    "pr_auc": {
      "labels": {"false": 0.997, "true": 0.445},
      "macro": 0.721, "micro": 0.978},
    "roc_auc": {
      "labels": {"false": 0.923, "true": 0.923},
      "macro": 0.923, "micro": 0.923},
    ...
  }
}
```


Figure 7: Result of https://ores.wikimedia.org/v3/scores/enwiki/?model_info&models=damaging

The output captured in Figure 7 shows a heavily trimmed JSON (human- and machine-readable) output of *model_info* for the "damaging" model in English Wikipedia. Note that many fields have been trimmed in the interest of space with an ellipsis ("..."). What remains gives a taste of what information is available. Specifically, there is structured data about what kind of model is being used, how it is parameterized, the computing environment used for training, the size of the train/test set, the basic set of fitness metrics, and a version number so that secondary caches know when to invalidate old scores. A developer using an ORES model in their tools can use these fitness metrics to make decisions about whether or not a model is appropriate and to report to users what fitness they might expect at a given confidence threshold.

Threshold optimization. When we first started developing ORES, we realized that operational concerns of Wikipedia's curators need to be translated into confidence thresholds for the prediction models. For example, counter-vandalism patrollers seek to catch all (or almost all) vandalism before it stays in Wikipedia for very long. That means they have an operational concern around the *recall* of a damage prediction model. They also like to review as few edits as possible in order to catch that vandalism. So they have an operational concern around the *filter rate*—the proportion of edits that are not flagged for review by the model[15].

By finding the threshold of prediction likelihood that optimizes the filter-rate at a high level of recall, we can provide vandal-fighters with an effective trade-off for supporting their work. We refer to these optimizations in ORES as *threshold optimizations* and ORES provides information about these thresholds in a machine-readable format so that tools can automatically detect the relevant thresholds for their wiki/model context.

Originally, when we developed ORES, we defined these threshold optimizations in our deployment configuration. But eventually, it became apparent that our users wanted to be able to search through fitness metrics to choose thresholds that matched their own operational concerns. Adding new optimizations and redeploying quickly became a burden on us and a delay for our users. In response, we developed a syntax for requesting an optimization from ORES in realtime using fitness statistics from the models tests. E.g. `maximum recall @ precision >= 0.9` gets a useful threshold for a counter-vandalism bot or `maximum filter_rate @ recall >= 0.75` gets a useful threshold for semi-automated edit review (with human judgement).


```
{"threshold": 0.30, ...,
 "filter_rate": 0.88, "fpr": 0.097,
 "precision": 0.21, "recall": 0.75}
```


Figure 8: Result of https://ores.wikimedia.org/v3/scores/enwiki/?models=damaging&model_info=statistics.thresholds.true.'maximumfilter_rate@recall⟩=0.75'

This result shows that, when a threshold is set on 0.299 likelihood of damaging=true, then you can expect to get a recall of 0.751, precision of 0.215, and a filter-rate of 0.88. While the precision is low, this threshold reduces the overall workload of vandal-fighters by 88% while still catching 75% of (the most egregious) damaging edits.

Explicit pipelines. We have designed the process of training and deploying ORES prediction models is to be repeatable and reviewable. Consider the following code that represents a common pattern from our model-building Makefiles:

Essentially, this code helps someone determine where the labeled data comes from (manually labeled via the Wiki Labels system). It makes it clear how features are extracted (using the `revscoring extract` utility and the `feature_lists.enwiki.damaging` feature set). Finally, this dataset of extracted features is used to cross-validate and train a model predicting the "damaging" label and a serialized version of that model is written to a file. A user could clone this repository, install the set of requirements, and run `make enwiki_models` and expect that all of the data-pipeline would be reproduced, and an exactly equivalent model obtained.

By explicitly using public resources and releasing our utilities and Makefile source code under an open license (MIT), we have essential implemented a turn-key process for replicating our model building and evaluation pipeline. A developer can review this pipeline for issues knowing that they are not missing a step of the process because all steps are captured in the Makefile. They can also build on the process (e.g. add new features) incrementally and restart the pipeline. In our own experience, this explicit pipeline is extremely useful for identifying the origin of our own model building bugs and for making incremental improvements to ORES' models.

At the very base of our Makefile, a user can run `make models` to rebuild all of the models of a certain type. We regularly perform this process ourselves to ensure that the Makefile is an accurate representation of the data flow pipeline. Performing complete rebuild is essential when a breaking change is made to one of our libraries. The resulting serialized models are saved to the source code repository so that a developer can review the history of any specific model and even experiment with generating scores using old model versions.

Received April 2018; revised July 2018; revised August 2018

```
datasets/enwiki.human_labeled_revisions.20k_2015.json:
	./utility fetch_labels \
		https://labels.wmflabs.org/campaigns/enwiki/4/ > $@

datasets/enwiki.labeled_revisions.w_cache.20k_2015.json: \
		datasets/enwiki.labeled_revisions.20k_2015.json
	cat $< | \
	revscoring extract \
		editquality.feature_lists.enwiki.damaging \
		--host https://en.wikipedia.org \
		--extractor $(max_extractors) \
		--verbose > $@

models/enwiki.damaging.gradient_boosting.model: \
		datasets/enwiki.labeled_revisions.w_cache.20k_2015.json
	cat $^ | \
	revscoring cv_train \
		revscoring.scoring.models.GradientBoosting \
		editquality.feature_lists.enwiki.damaging \
		damaging \
		--version=$(damaging_major_minor).0 \
		(... model parameters ...)
		--center --scale > $@
```

Figure 9: Makefile rules for the English damage detection model from https://github.com/wiki-ai/editquality