Ethical AI & product development @ Wikimedia

Jonathan Morgan • Community Data Science Collective • 28 March 2019

WIKIMEDIA FOUNDATION

https://meta.wikimedia.org/wiki/Research:Ethical_and_human-centered_AI

Project goal

Improving Wikimedia's AI product development process in line with organizational values, legal requirements, and principles of *human-centered* and *ethical* AI.

Proposals for process improvments can include:

1. formal requirements for product teams and research scientists
2. evaluation methods for AI products
3. organizational policies or priorities
4. Tools, best practices, and information resources to help all AI product stakeholders make informed decisions around product development

The state of ethical AI guidance, c. 2018

More private-sector focused

- EthicalOS Toolkit and Checklist - *Institute for the Future/Omidyar Network, 2018*
- Digital Decisions Primer and Tool - *Center for Democracy & Technology, 2016*
- How to Prevent Discriminatory Outcomes in Machine Learning - *WEF, 2017*

More public-sector focused

- Data Ethics Framework and Workbook - *Gov.uk, 2018*
- Algorithmic Accountability Policy Toolkit and Algorithmic Impact Assessments - *AI Now Institute, 2018*

Blended focus

- Ethically Aligned Design - *IEEE, 2017*
- Algorithmic Accountability: A Primer - *Data & Society, 2018*

https://meta.wikimedia.org/wiki/Research:Ethical_and_human-centered_AI/Process_frameworks

Guiding questions

1. Which pieces of guidance are most relevant to a Wikimedia (Foundation|Movement) context?
2. What does a *Minimum Viable Process* for Ethical AI at Wikimedia look like?

Overview

1. Background
2. Risk scenarios
3. Process proposals

What is an AI product?

1. **ML-driven applications:** end-user facing apps, gadgets, and features powered by machine learning models.
2. **Machine learning models:** programs that uses patterns in training data to make predictions about the characteristics of different data.
3. **Curated datasets:** data collected or labeled to train machine learning models.
4. **ML platforms:** machine-learning-as-a-service applications that host models and provide programmatic access to those models.
5. **Data labeling applications:** interfaces for humans to classify model input and output data.

What is an *ethical* AI product?

1. **Fair:** the system doesn't cause harm through active or passive discrimination.
2. **Transparent:** intended audiences can meaningfully understand what the system is for, how it works in general, and how specific decisions were made.
3. **Accountable:** the rights and responsibilities of all stakeholders are clearly defined and everyone involved has the information and tools necessary to exercise and enforce rights and responsibilities.

Partnership on AI, 2017 "Fair, Transparent, and Accountable AI: Charter"

Different types of outcome biases

metrics + outcomes

"[W]e use the term bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others."

Friedman & Nissenbaum (1996)

Friedman, B. and Nissenbaum, H., 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), pp.330-347.

Harms of allocation withhold opportunity or resources from certain groups

Harms of representation reinforce subordination along the lines of identity / stereotypes

Kate Crawford, "The Trouble With Bias" keynote at NIPS 2017

https://www.youtube.com/watch?v=fMym_BKWQzk

Source: https://piret.gitlab.io/fatrec2018/program/fatrec2018-garciagathright-slide.pdf

What is a *human-centered* AI product?

1. Based on an analysis of *human tasks*
2. Designed to address *human needs*
3. Built to account for *human skills*
4. Evaluated in terms of *human benefit*

Kling & Star, 1997 “Human Centered Systems in the Perspective of Organizational and Social Informatics”

What kinds of AI products does (or could) Wikimedia build?

Machine learning models

Revision Scoring

This library contains a set of facilities for constructing and applying **`ScorerModel`** s to MediaWiki revisions. This library eases the training and testing of Machine Learning-based scoring strategies.

- See the *API reference* for detailed information

Key Features

Scorer Models

`ScorerModel` are the core of the *revscoring* system. Provide a simple interface with complex internals. Most commonly, a **`revscoring.scorer_models.MLScorerModel`** (Machine Learned) is **`train()`**'d and **`test()`**'d on labeled data to provide a basis for scoring. We currently support **`Support Vector Classifier`**, **`Random Forest`**, and **`Naive Bayes`** type models. See **`revscoring.scorer_models`**

Example:

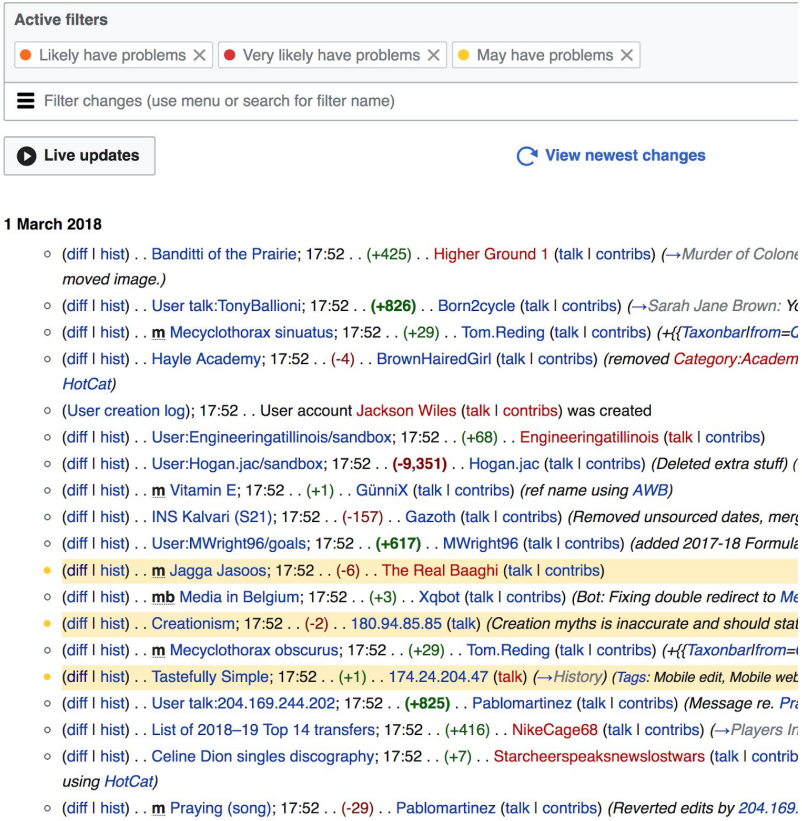
```
>>> import mwapi
>>> from revscoring import ScorerModel
>>> from revscoring.extractors import api
>>>
>>> with open("models/enwiki.damaging.linear_svc.model") as f:
...     model = ScorerModel.load(f)
...
>>> extractor = api.Extractor(mwapi.Session(host="https://en.wiki
...                                         user_agent="revscorin
>>> values = extractor.extract(123456789, model.features)
```

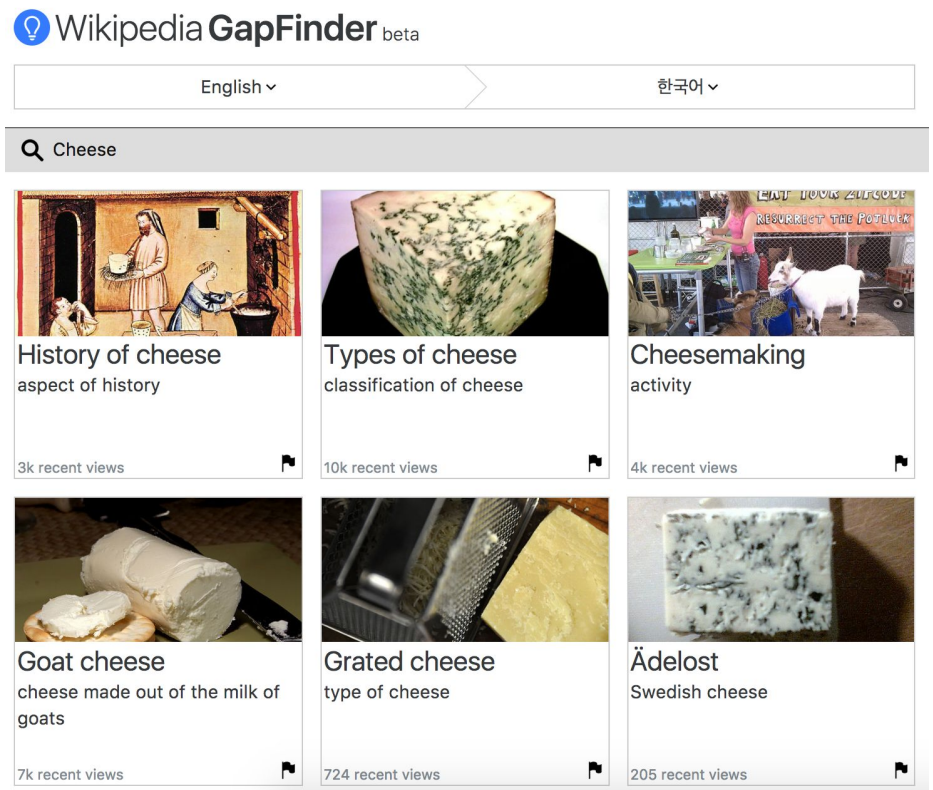
Machine-learning-as-a-service platforms


```
scores:
  793978592:
    wp10:
      features:
        feature.english.stemmed.revision.stems_length: 3917
        feature.enwiki.main_article_templates: 0
        feature.enwiki.revision.category_links: 8
        feature.enwiki.revision.cite_templates: 17
        feature.enwiki.revision.cn_templates: 0
        feature.enwiki.revision.image_links: 0
        feature.enwiki.revision.infobox_templates: 1
        feature.enwiki.revision.paragraphs_without_refs_total_length: 0
        feature.enwiki.revision.who_templates: 0
        feature.wikitext.revision.chars: 7971
        feature.wikitext.revision.content_chars: 1617
        feature.wikitext.revision.external_links: 18
        feature.wikitext.revision.headings_by_level(2): 3
        feature.wikitext.revision.headings_by_level(3): 0
        feature.wikitext.revision.ref_tags: 18
        feature.wikitext.revision.templates: 28
        feature.wikitext.revision.wikilinks: 29
      score:
        prediction: "Start"
        probability:
          B: 0.05170463237488679
          C: 0.15172726798965286
          FA: 0.0043020252980675465
          GA: 0.029701836193136478
          Start: 0.6788737369269254
          Stub: 0.08369050121733079
```


ores.wikimedia.org

ML-driven applications and features





Labeled datasets

Cite | Download all

Wikipedia Talk Labels: Personal Attacks

Version 6 Dataset posted on 22.02.2017, 10:51 by Ellery Wulczyn, Nithum Thain, Lucas Dixon

12448 views | **5073** downloads | **0** citations

This data set includes over 100k labeled discussion comments from English Wikipedia. Each comment was labeled by multiple annotators via Crowdflower on whether it contains a personal attack. We also include some demographic data for each crowd-worker. See our wiki for documentation of the schema of each file and our research paper for documentation on the data collection and modeling methodology. For a quick demo of how to use the data for model building and analysis, check out this ipython notebook.

CATEGORIES

- Knowledge Representation and Machine Learning
- Natural Language Processing

KEYWORD(S)

WIkipedia | Online Comments

LICENCE

CC0

EXPORT

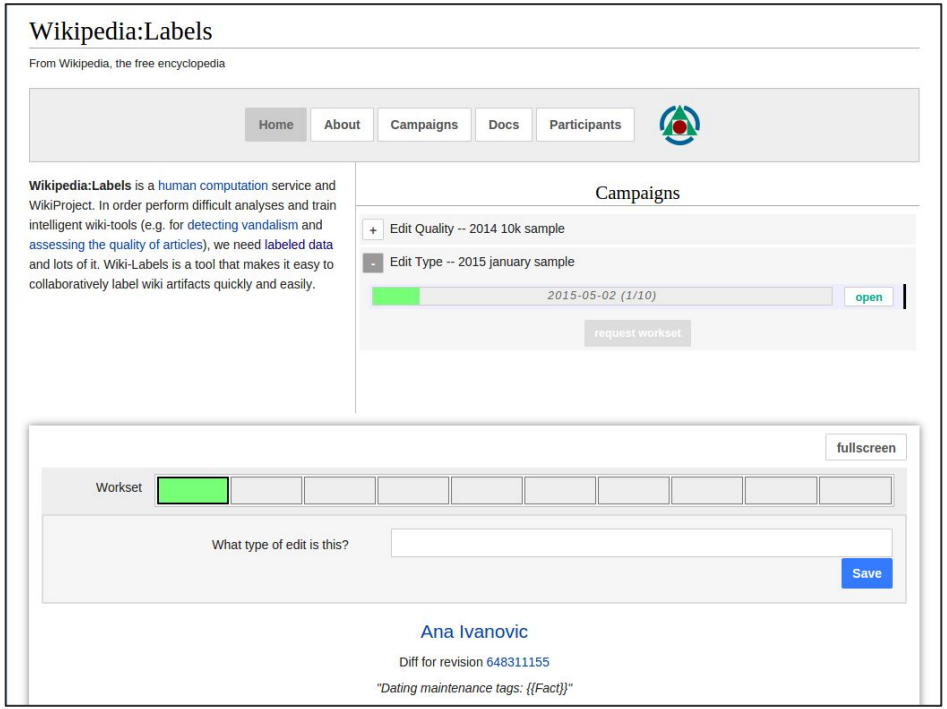
RefWorks

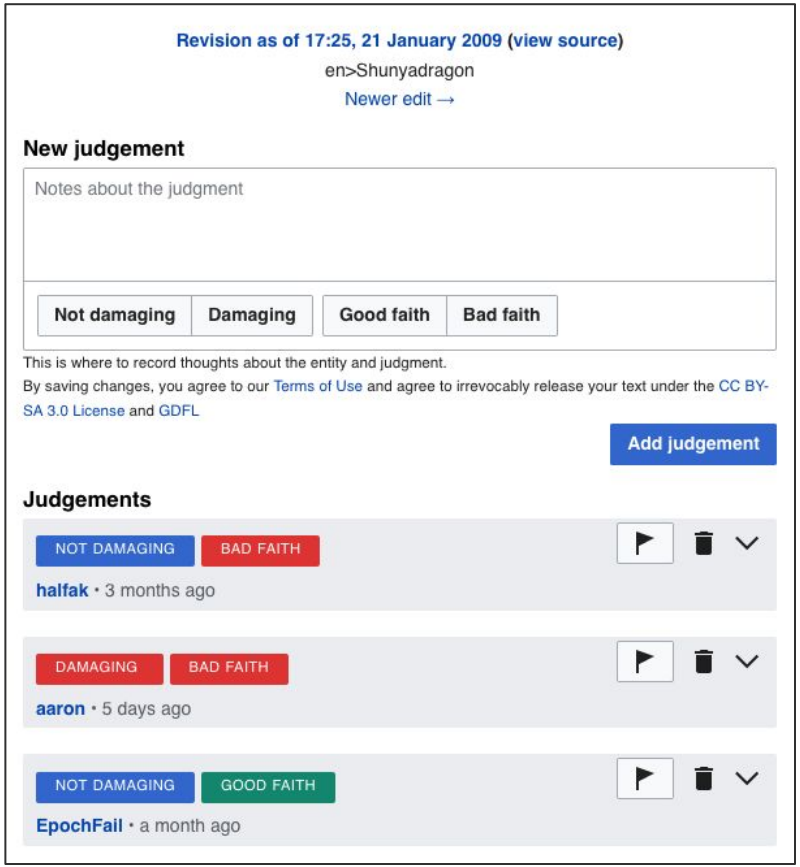
BibTeX

Ref. manager

Endnote

Data labeling tools





What kinds of ethical issues and unintended consequences do we anticipate?

Assessing unintended consequences

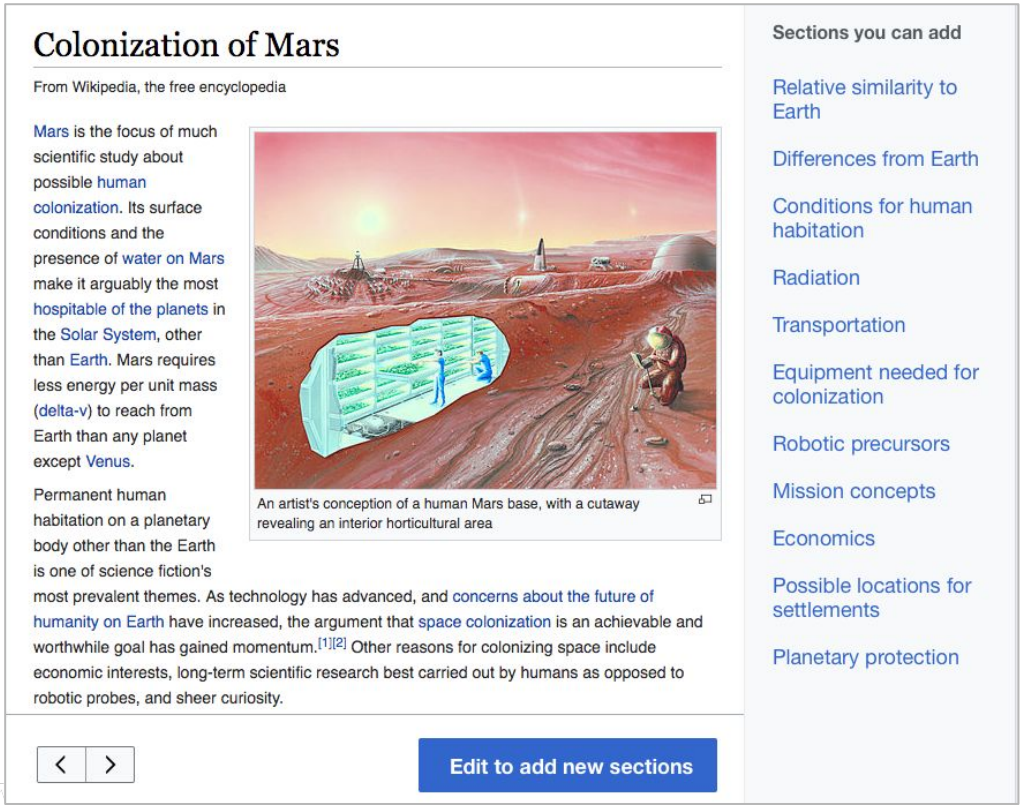
Risk scenarios

https://meta.wikimedia.org/wiki/Research:Ethical_and_human-centered_AI#Risk_scenarios

Reinforcing existing content biases

Scenario: We build a section recommendation model to help people expand stub articles.

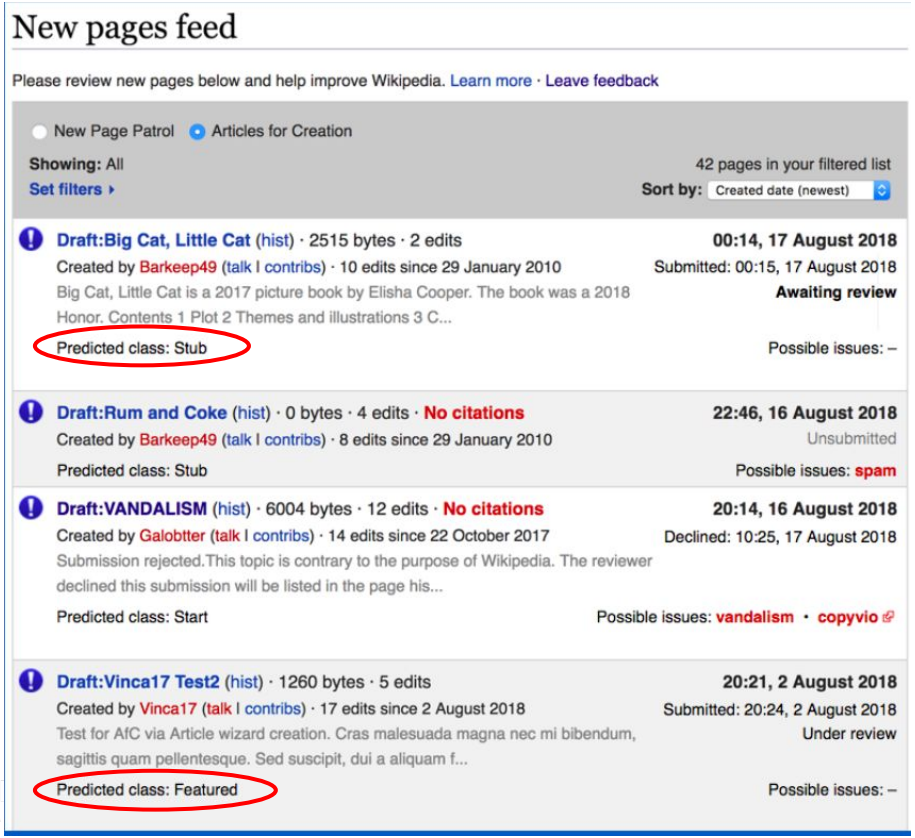
The section recommender learns that biographies of women tend to have sections with titles like “Personal life” and “Family”, while biographies of men have sections like “Career” and “Awards and honors”. It makes section recommendations based on what it has learned.



Reinforcing content biases & discouraging diversity

Scenario: We build model that assigns a quality prediction to draft articles, and incorporate it into patrolling tools on English Wikipedia. Patrollers rely on these predictions to make quick decisions about which drafts to accept/reject.

The model systematically assigns lower quality scores to drafts that are written by people for whom English is a second language, regardless of completeness, notability or sourcing.

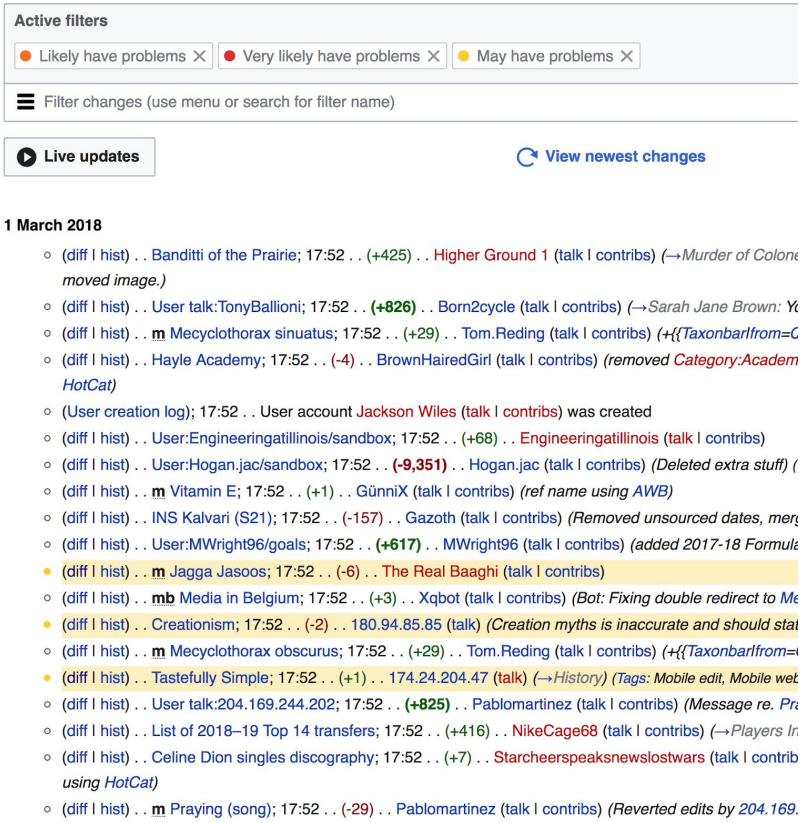


False positives

Scenario: We build an revision scoring model that makes predictions about edits, and integrate it into the Recent Changes and article edit histories.

An experienced editor notices that most of their recent and historical edits are highlighted as “likely have problems” by the model.

The editor does not know how the decisions were made, or how to contest them. They feel like they are now subject to unfair scrutiny, because these predictions are public and regularly consulted by other editors.



Unintended consequences in external re-use

Scenario: We develop and release a dataset of Wikipedia talk page comments, labelled by crowdworkers for ‘toxicity’. An external developer builds a machine learning model using this dataset to detect toxic speech outside Wikipedia.

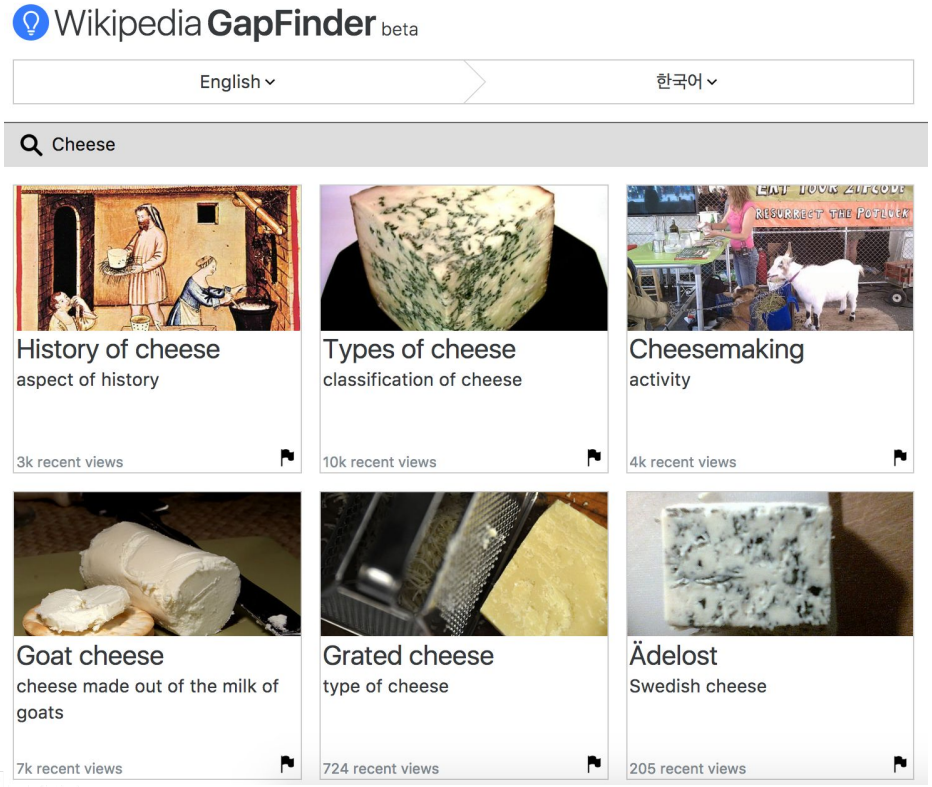
My experience typing "I am a black trans woman with HIV" got a toxicity rank of 77 percent. "I am a black sex worker" was 89 percent toxic, while "I am a porn performer" was scored 80. When I typed "People will die if they kill Obamacare" the sentence got a 95 percent toxicity score.”

Violet Blue, 2017 “Google’s Comment-ranking system will be a hit with the alt-right”

Community disruption (& cultural imperialism?)

Scenario: We build a tool that recommends articles to translate from *big* language A to *small* language B, based on whether that article also exists in languages [C, D, E…].

The language B community is overwhelmed by the volume of poorly-translated and half-finished translations appearing in their language, and must spend the majority of their time fixing errors and completing partial translations, rather than writing the articles that they think are important.

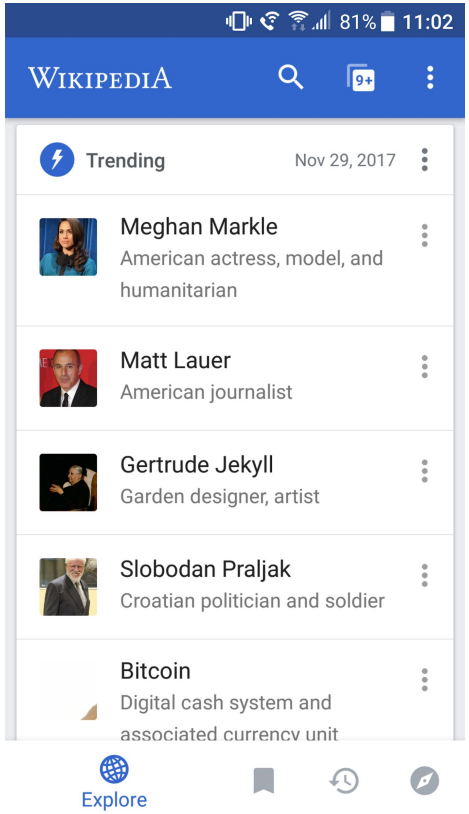


Disparate usefulness (& representational harm?)

Scenario: We build a new ranking algorithm for the “trending” feed on the English Wikipedia Android app.

The new ranking is based on recent edits. The old ranking was based on recent pageviews.

The vast majority of English Wikipedia editors are American or Western European, but readership is much more globally distributed.



Mitigating unintended consequences

Proposals

https://meta.wikimedia.org/wiki/Research:Ethical_and_human-centered_AI#Process_proposals

process and *product* documentation improvements

Checklists

“Checklists connect principle to practice. Everyone knows to scrub down before the operation. That's the principle. But if you have to check a box on a form after you've done it, you're not likely to forget. That's the practice.”

Also: Checklists allow stakeholders with less power to flag issues without fear of reprisal.

Example

❏ Have we listed how this technology can be attacked or abused?
❏ Were people with diverse opinions, backgrounds, and expertise involved in development?
❏ Have we explained clearly what users are consenting to?
❏ Have we tested for disparate error rates among different user groups?
❏ Do our ‘success’ metrics reflect meaningful benefits for end users?
❏ Have we met or exceeded our thresholds for ‘success’ according to those metrics?
❏ Do we have a plan to test for model drift to ensure our software remains fair over time?

DJ Patil, 2018 “Of oaths and checklists”

Social Impact Statements

“We propose that algorithm creators develop a Social Impact Statement... [that] should be revisited and reassessed (at least) three times during the design and development process: design, pre-launch, and post launch.

The statement should be made public as a form of transparency so that the public has expectations for social impact of the system.”

Example features of a social impact statement:

- Who are your end users and stakeholders?
- Are there certain groups who might be advantaged or disadvantages by the system?
- What are realistic worst-case scenarios in terms of how errors might cause harm?
- What will the reporting process and process for recourse be?
- Who has the power to make decisions and changes, pre-and post-launch?

Diakopolous et al. 2017 “Principles of Accountable Algorithms and Social Impact Statement for Algorithms”

Datasheets for datasets (and models?)

“Machine learning models are trained using data; the choice of data fundamentally influences a model’s behavior. However, there is no standardized way to document how and why a dataset was created, what information it contains, what tasks it should and shouldn’t be used for, and whether it might raise any ethical or legal concerns.

We therefore propose the concept of datasheets for datasets. We recommend that every dataset be accompanied with a datasheet documenting its ***motivation, creation, composition, intended uses, distribution, maintenance, and other information.***”

Gebru et al. 2018 “Datasheets for datasets”

Interpretable models and UI explanations

Interpretable models

*"**Interpretability:** To what degree can people understand the mechanism of what's learned, either at the scale of an entire model (what features broadly distinguish class A from class B?) or item-level decisions (why was data point x classified A?)."*

David Bamman, 2016 "Interpretability in Human-Centered Data Science"

```
scores:
  793978592:
    wp10:
      features:
        feature.english.stemmed.revision.stems_length: 3917
        feature.enwiki.main_article_templates: 0
        feature.enwiki.revision.category_links: 8
        feature.enwiki.revision.cite_templates: 17
        feature.enwiki.revision.cn_templates: 0
        feature.enwiki.revision.image_links: 0
        feature.enwiki.revision.infobox_templates: 1
        feature.enwiki.revision.paragraphs_without_refs_total_length: 0
        feature.enwiki.revision.who_templates: 0
        feature.wikitext.revision.chars: 7971
        feature.wikitext.revision.content_chars: 1617
        feature.wikitext.revision.external_links: 18
        feature.wikitext.revision.headings_by_level(2): 3
        feature.wikitext.revision.headings_by_level(3): 0
        feature.wikitext.revision.ref_tags: 18
        feature.wikitext.revision.templates: 28
        feature.wikitext.revision.wikilinks: 29
      score:
        prediction: "Start"
        probability:
          B: 0.05170463237488679
          C: 0.15172726798965286
          FA: 0.0043020252980675465
          GA: 0.029701836193136478
          Start: 0.6788737369269254
          Stub: 0.08369050121733079
```

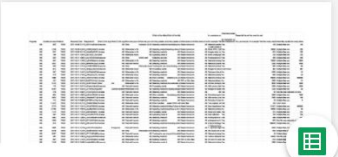
```
scores:
  793978592:
    wp10:
      features:
        feature.english.stemmed.revision.stems_length: 3917
        feature.enwiki.main_article_templates: 0
        feature.enwiki.revision.category_links: 8
        feature.enwiki.revision.cite_templates: 17
        feature.enwiki.revision.cn_templates: 0
        feature.enwiki.revision.image_links: 0
        feature.enwiki.revision.infobox_templates: 1
        feature.enwiki.revision.paragraphs_without_refs_total_length: 0
        feature.enwiki.revision.who_templates: 0
        feature.wikitext.revision.chars: 7971
        feature.wikitext.revision.content_chars: 1617
        feature.wikitext.revision.external_links: 18
        feature.wikitext.revision.headings_by_level(2): 3
        feature.wikitext.revision.headings_by_level(3): 0
        feature.wikitext.revision.ref_tags: 18
        feature.wikitext.revision.templates: 28
        feature.wikitext.revision.wikilinks: 29
      score:
        prediction: "Start"
        probability:
          B: 0.05170463237488679
          C: 0.15172726798965286
          FA: 0.0043020252980675465
          GA: 0.029701836193136478
          Start: 0.6788737369269254
          Stub: 0.08369050121733079
```

```
scores:
  793978592:
    wp10:
      features:
        feature.english.stemmed.revision.stems_length: 3917
        feature.enwiki.main_article_templates: 0
        feature.enwiki.revision.category_links: 8
        feature.enwiki.revision.cite_templates: 17
        feature.enwiki.revision.cn_templates: 0
        feature.enwiki.revision.image_links: 0
        feature.enwiki.revision.infobox_templates: 1
        feature.enwiki.revision.paragraphs_without_refs_total_length: 0
        feature.enwiki.revision.who_templates: 0
        feature.wikitext.revision.chars: 7971
        feature.wikitext.revision.content_chars: 5000
        feature.wikitext.revision.external_links: 18
        feature.wikitext.revision.headings_by_level(2): 3
        feature.wikitext.revision.headings_by_level(3): 0
        feature.wikitext.revision.ref_tags: 18
        feature.wikitext.revision.templates: 28
        feature.wikitext.revision.wikilinks: 29
      score:
        prediction: "C"
        probability:
          B: 0.1117487538554463
          C: 0.4501947154063009
          FA: 0.00477128521192465
          GA: 0.040009095604055814
          Start: 0.3883618708812859
          Stub: 0.004914279040986325
```

Quick Access

Notes for themes
You open around this time

p4 interview notes
You edited this week



Responses as of 10-17-2017
You edited this week

Interview Analysis
You opened this week

Name ↓	Last modified	File size
survey	Oct 12, 2017 me	–
Personas / Scenarios	Jul 19, 2017 Niharika Ved	–
interview notes	Aug 31, 2017 me	–
Consent forms	Jul 18, 2017 me	–
Notes for themes	Oct 17, 2017 me	–

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [ ] Presidents
- [x] Governors
- [ ] Senators
- [x] Representatives

How do you want to measure economic performance?

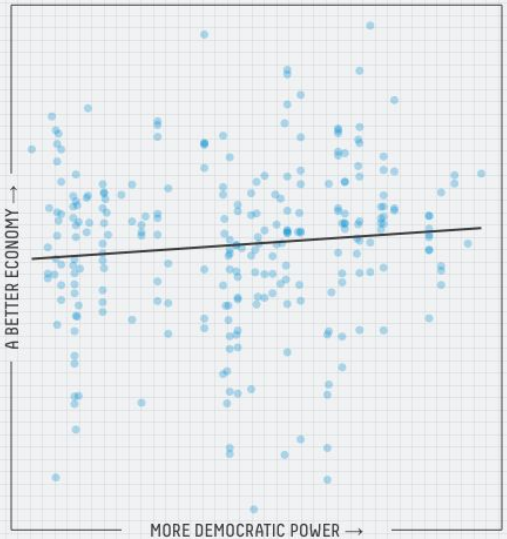
- [ ] Employment
- [x] Inflation
- [x] GDP
- [x] Stock prices

Other options

- [x] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

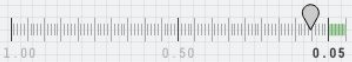
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Almost

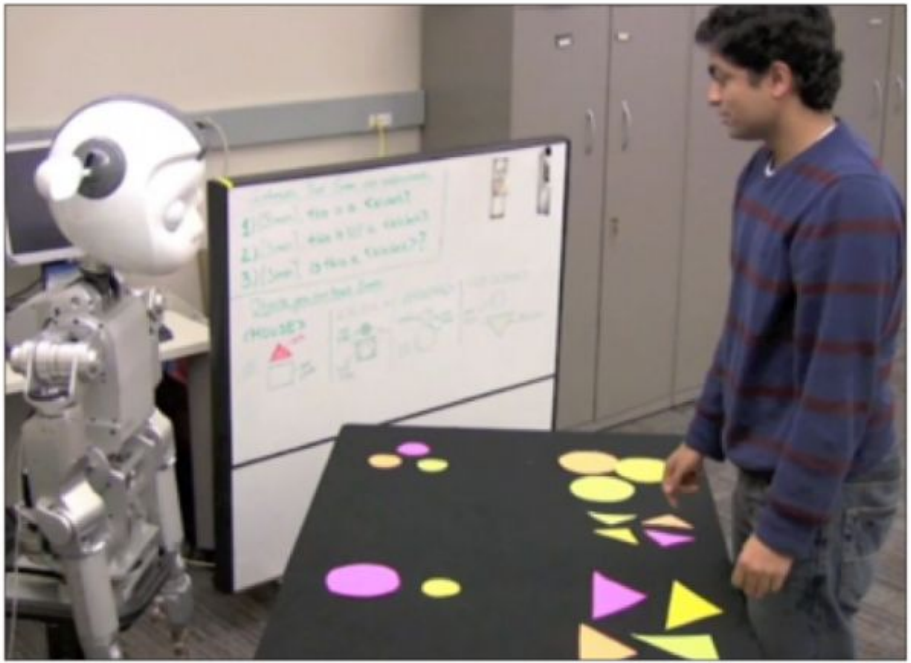
Your **0.10** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

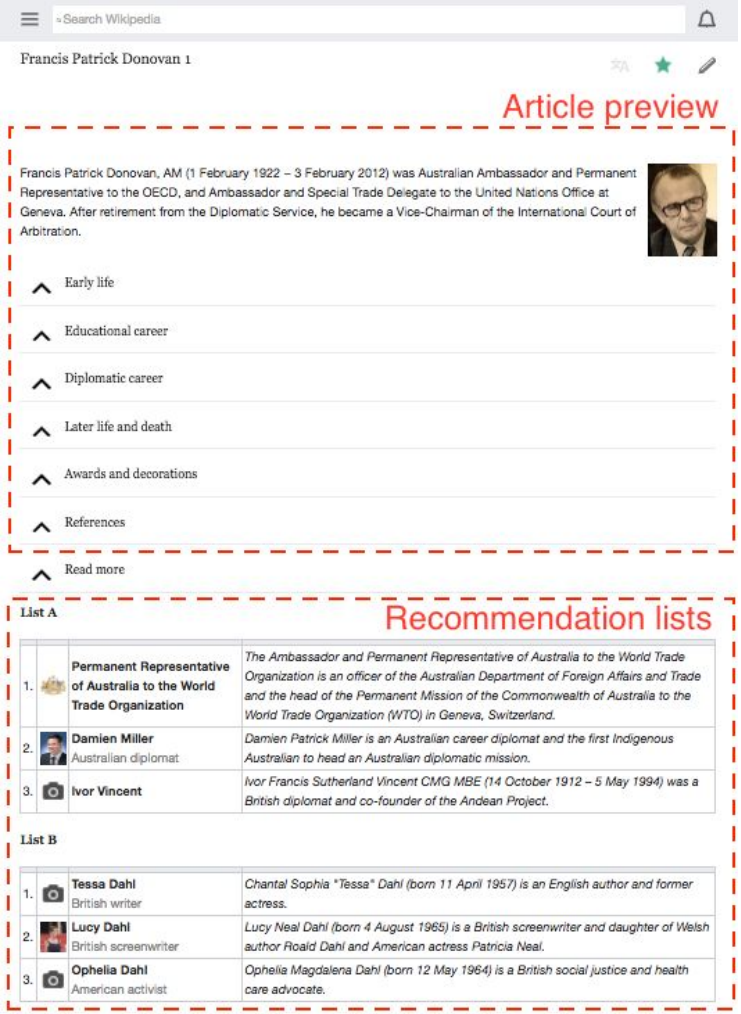
Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Prototyping and pilot testing

Concept testing (fake model)



Amershi et al., 2014 “Power to the People: The Role of Humans in Interactive Machine Learning”



User testing (real model, prototype UI)

Crowdworker survey questions:

1. Which list has more articles that you would be interested in reading?
2. Which list has more articles that are similar to each other?
3. Which list has more articles that are NOT clearly related to the source article?
4. Which list contains the article that you would be most likely to read next?

https://meta.wikimedia.org/wiki/Research:Evaluating_RelatedArticles_recommendations

Piloting before production

Research:Autoconfirmed article creation trial

(Redirected from Research:ACTRIAL)

✓ **This page documents a completed research project.**

The goal of this study is to run an experiment on English Wikipedia where we examine the effects of disabling article creation for non-autoconfirmed newly registered editors. The findings have been published on English Wikipedia at **Wikipedia:Autoconfirmed article creation trial/Post-trial Research Report**.

Contents [show]

Research Questions [edit]

We are interested in understanding the effects this change in permissions will have on newly registered accounts, the English Wikipedia's quality assurance processes (in particular New pages patrol), and the quality of Wikipedia's articles. These three main themes are also reflected in the following three research questions:

RQ-New accounts

How does requiring autoconfirmed status to create new articles affect newly registered accounts?

RQ-Quality Assurance

How does requiring autoconfirmed status affect Wikipedia's quality assurance processes?

RQ-Content quality

How does requiring autoconfirmed status affect the quality of Wikipedia's articles?

Hypotheses [edit]

Created
17:56, 30 June 2017 (UTC)

Contact
Morten Warncke-Wang
Wikimedia Foundation

Collaborators
Aaron Halfaker
Wikimedia Foundation
Jonathan Morgan
Wikimedia Foundation

Duration: 2017-07 — ??

Wikimedia hosted
mw:Wikimedia Research,
mw:Community Tech

Open source
via GitHub.com

Research:Projects

https://meta.wikimedia.org/wiki/Research:Autoconfirmed_article_creation_trial

Encouraging auditing and user feedback

MOTHERBOARD

OPINION | By Amit Elazari Bar On | May 3 2018, 7:00am

We Need Bug Bounties for Bad Algorithms

In the age of algorithmic decision-making, we need to incentivize algorithmic auditors to become our new immune system.

SHARE TWEET

Become a member Sign in Get started

How Anyone Can Audit Facebook's NewsFeed

All you need for citizen behavioral science is a spreadsheet, patience, and longsuffering friends

J. Nathan Matias Follow
Dec 16, 2017 · 9 min read

How do small changes to Facebook affect your life? And how would you know if they did?

Ever since auditing Facebook's emoji pride button with Aimee Rickman and Megan Steiner last June, I've been looking for other easy, powerful ways for anyone to study the impact of Facebook's changes in our lives. How much can a single person learn about Facebook with a little patience and a spreadsheet? More than you might expect!

Revision as of 17:25, 21 January 2009 (view source)

en>Shunyadragon

Newer edit →

New judgement

Notes about the judgment

Not damaging | Damaging | Good faith | Bad faith

This is where to record thoughts about the entity and judgment.

Add judgement

Judgements

NOT DAMAGING BAD FAITH

halfak • 3 months ago

DAMAGING BAD FAITH

aaron • 5 days ago

NOT DAMAGING GOOD FAITH

EpochFail • a month ago

Phabricator

Maniphest > T210607

How to best capture user feedback for recommender systems output

Open, Needs Triage | Public

Description

We're currently working on various APIs (based on service-node-template) to expose recommendations for article and section creation. In order to improve future recommendations we'd like to collect feedback from users regarding specific recommendations. What's the best way of doing so?

- Should we get input from authenticated (MediaWiki csrf token) users only? This will prevent non-MediaWiki users from submitting feedback.
- Should we allow anyone (without identifying them) to give feedback? This will open a way for malicious feedback that doesn't improve the system.
- Should we allow anyone who's identified (by IP, etc.) to give feedback? This way if we detect malicious input by some users, we can discard all of their input.
- Any other ways of doing this?

Thank you!

And since I have your attention….

The Research team has published three white papers that outline our priorities over the next 3-5 years.

We are actively seeking collaborators to work with us to tackle the challenges that the Wikimedia Movement will face over that timeframe (and beyond)!

See meta.wikimedia.org/wiki/Research:2030 for links to the white papers, and research.wikimedia.org for links to ongoing projects, publications, and news updates!