Social Computing Capstone

Day 18: Designing AI for Communities

CSE 481p | Winter 2022

Ruotong Wang
PhD Student | University of Washington, Allen School of Computer Science & Engineering

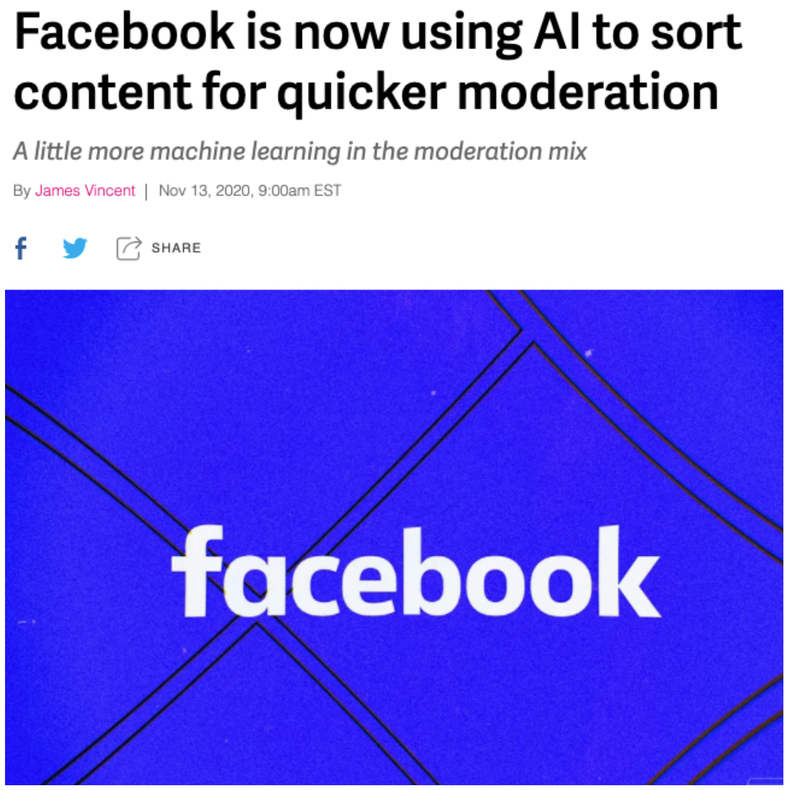
Schedule for today’s class

- Lecture on today’s topic (15 min)
- Group working time (65 min)

Recall from previous lectures…

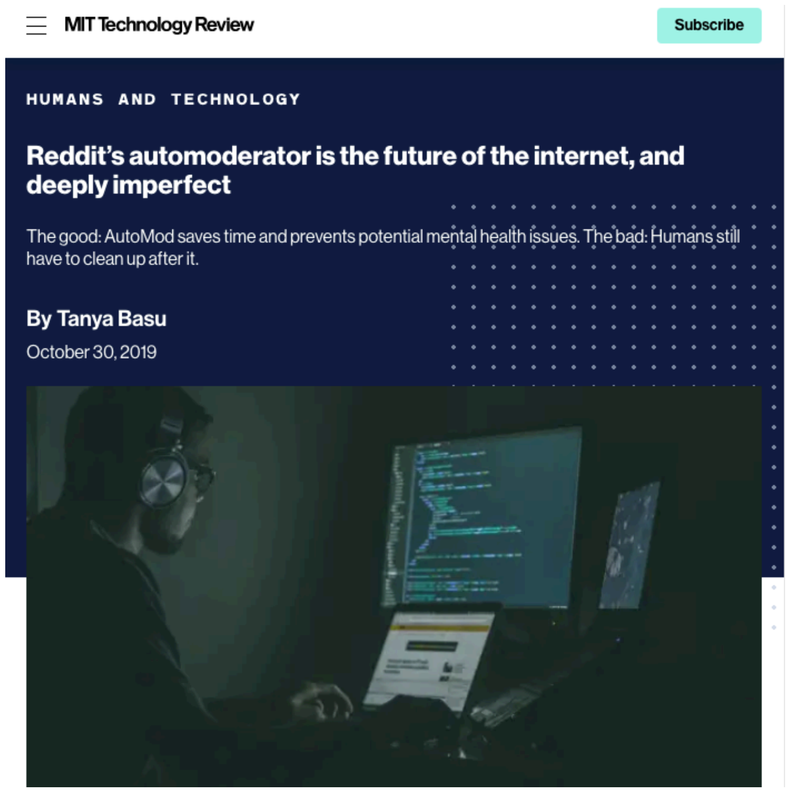
- **Commercial content moderation:** thousands of paid contractors who work for the platform reviewing content
- **Community content moderation:** members of the community, or moderators who run the community, handle reports and proactively remove comments
- **Algorithmic content moderation:** AI systems trained on previously removed comments predict whether new comments should be removed

Algorithmic content moderation



Facebook

Wikipedia ORES



Reddit Automod

What does "good" mean for AI systems in online communities?

Designing AI systems for communities is hard.
These algorithms could fail...
even when they are doing their best

These algorithms could fail... when facing edge cases

WEB TECH TUMBLR

Tumblr is already flagging innocent posts as porn

26

TECH YOUTUBE CULTURE

YouTube is still restricting and demonetizing LGBT videos — and adding anti-LGBT ads to some

Algorithms will always face cases that are **at the margin**: outside the situations seen in their training data.

the algorithm has to generalize: to "fill in the gaps" between the policy implied by training data and a new case the likes of which it has never seen before.

Algorithms **can not reflexively refine** their decision criteria as they reason through a novel situation. They at best refine their criteria only after the decision is made.

[Alkhatib and Bernstein 2019]

These algorithms could fail… when they can't fit in the social context

Article

American Behavioral Scientist
57(5) 664–688
© 2012 SAGE Publications
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0002764212469365
abs.sagepub.com
SAGE

The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline

Aaron Halfaker[1], R. Stuart Geiger[2], Jonathan T. Morgan[3], and John Riedl[1]

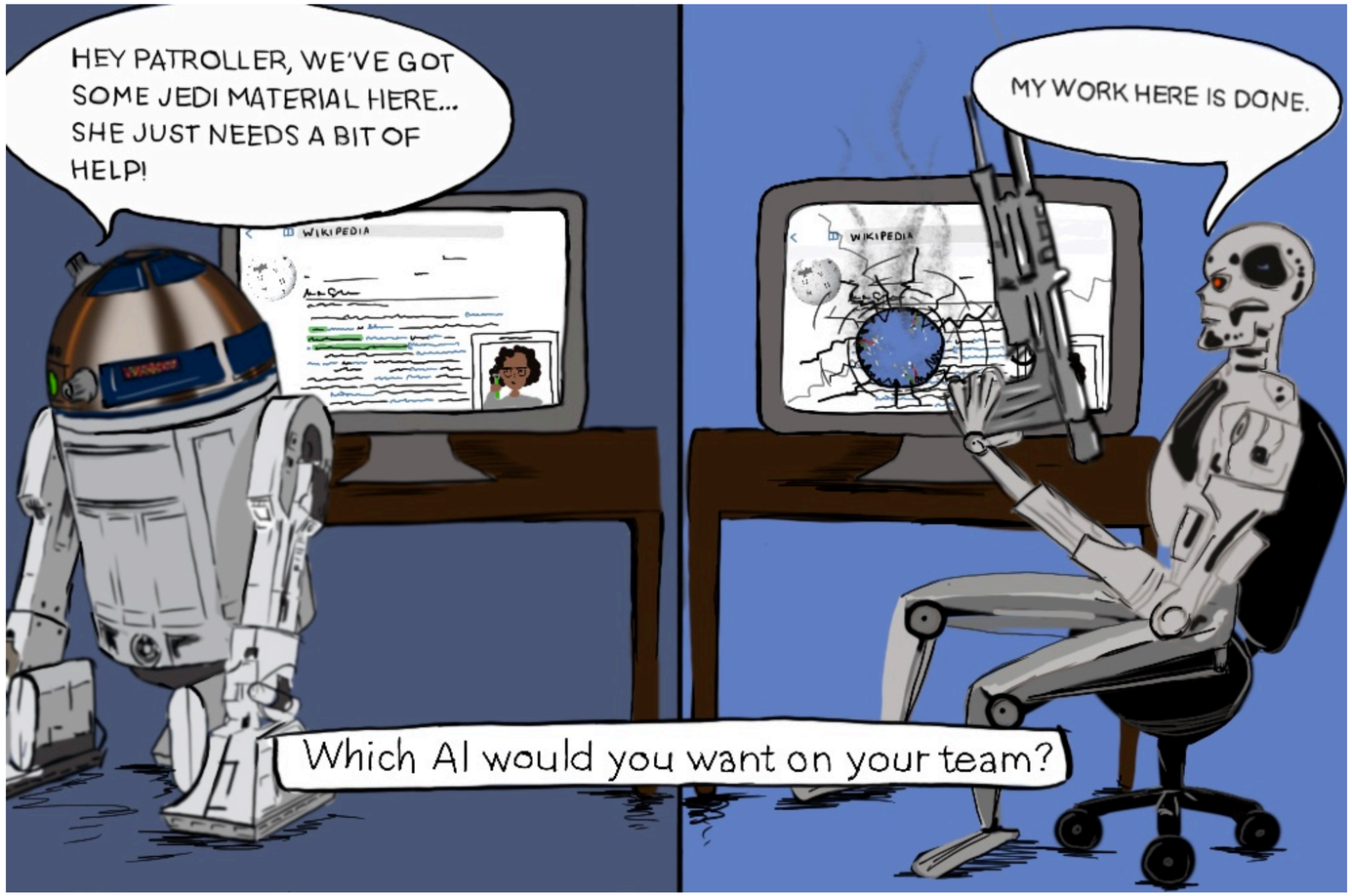
Abstract

Open collaboration systems, such as Wikipedia, need to maintain a pool of volunteer contributors to remain relevant. Wikipedia was created through a tremendous number of contributions by millions of contributors. However, recent research has shown that the number of active contributors in Wikipedia has been declining steadily for years and suggests that a sharp decline in the retention of newcomers is the cause. This article presents data that show how several changes the Wikipedia community made to manage quality and consistency in the face of a massive growth in participation have ironically crippled the very growth they were designed to manage. Specifically, the restrictiveness of the encyclopedia's primary quality control mechanism and the algorithmic tools used to reject contributions are implicated as key causes of decreased newcomer retention. Furthermore, the community's formal mechanisms for norm articulation are shown to have calcified against changes—especially changes proposed by newer editors.

AI tools failed because they are insensitive to **contributors' motivations** and **community values**.

[Halfaker et al. 2012]

These algorithms could fail... when they can't fit in the social context



Original Artwork contributed by: Laura Clapper

[Smith et al. 2020]

How to design AI systems for communities?

Short answer: Keep community in the loop

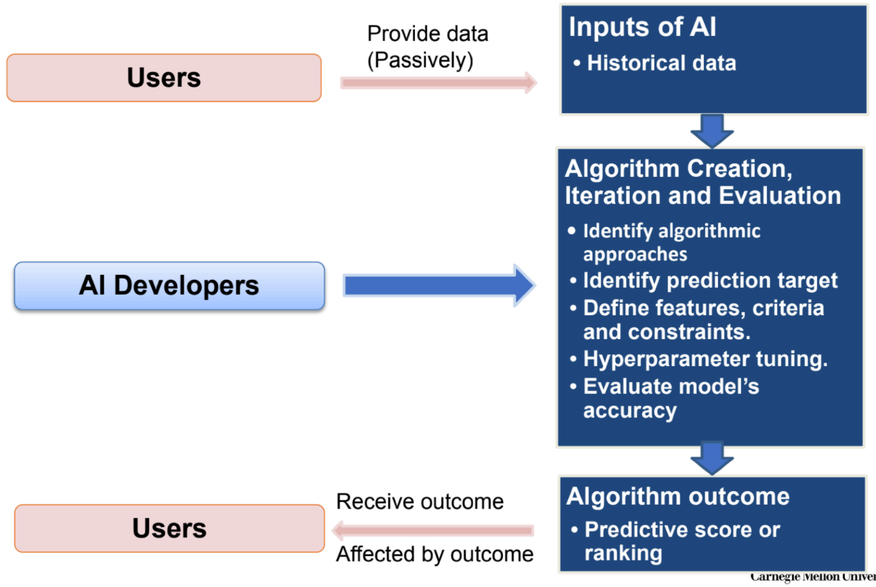
Principle 1: Understand community stakeholders' values related to algorithms

- **Stakeholders:** one who is involved in or affected by a course of action [Merriam-Webster]
- **Values:** what a person or group of people consider important in life [Borning and Muller 2012]

Principle 1: Understand community stakeholders' values related to algorithms

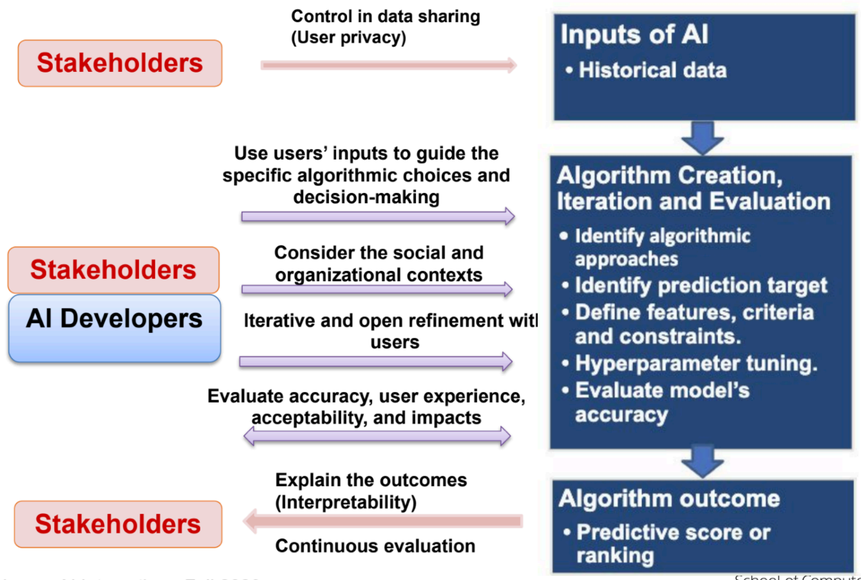
- **Stakeholders:** one who is involved in or affected by a course of action [Merriam-Webster]
- **Values:** what a person or group of people consider important in life [Borning and Muller 2012]

Traditional development pipeline



Carnegie Mellon University
School of Computer Science

Human-centric approach



sity
School of Computer Science

[Zhu et al. 2018]

Principle 1: Understand community stakeholders' values related to algorithms

- **Stakeholders:** Facebook users in different countries
- **Values:** severity of harmful content

- Values are not universal

Harmful Content	*Δrank*
Minor Sexualization	45
Self Injury Depiction	45
Adult Sexual Activity	43
Regulated Goods: Marijuana Sale	43
Sexually Explicit Language	43
Regulated Goods: Endangered Species Sale	41
Graphic Violence: Mutilated Humans	39
Interrupting Platform Services	39
Voter Fraud	39
Sexual Solicitation	39
Criminal Group Coordination	38
Criminal Group Propaganda	38
Eating Disorder Promotion	38
Celebrating Crime	37
Graphic Violence: Animal Abuse	36
Regulated Goods: Firearm Sale	36
Graphic Violence: Child Abuse	35
Suicide Depiction	34

https://doi.org/10.1371/journal.pone.0256762.t005

Types of harmful content that had max ranking differences (Δ*rank*) of at least 33, or half of the total number of rank positions. [Jiang et al. 2021]

Principle 2: Incorporate and balance these values into algorithms

Convergent Community Values for Machine Learning Systems on Wikipedia

	Effort Reduction	Human Authority	Workflow Support	Positive Engagement	Community Trust
ML systems should...	*...reduce the effort of community maintenance*	*...maintain human judgement as the final authority*	*...support differing peoples' differing workflows*	*...encourage positive engagement w/ diverse editors*	*...establish the trustworthiness of people & algorithms*

[Smith et al. 2020]

So, maybe when ORES detects damaging (but good faith) edits in Recent Changes, those edits could receive special treatment. For example, ... direct a patroller to first reach out to you before reverting, provide some scaffolded text like, "*Hi @yourhandle! Thanks for making your first edit to Wikipedia! Unfortunately, our algorithm detected an issue... It seems like you meant well, so I wanted to see if you could fix this by adding a citation so that I don't have to revert it?*" [Smith et al. 2020]

Principle 3: Iteratively evaluate algorithms based not only on accuracy, but also on their acceptability and broader impacts

- Fairness
- Trust
- Transparency
- Level of human engagement
- Impact on community norm
- …

Complicated decisions by algorithms in other communities…

- Work assignment
- Resource allocation
- Hiring
- Loan/credit assessment
- Child Maltreatment
- …

Summary

- Designing AI systems for communities is hard.
- We should keep community in the loop when designing algorithms for communities

Tips for posters

- Think about the core message that you want to send via the poster.
- Make sure what’s interesting is what stands out
- Minimizes text
- Use high-quality images
- Use consistent color scheme
- Send PDF to Ruotong by Wednesday 3pm

Group work time!