Keeping Community in the Machine-Learning Loop

C. Estelle Smith
PhD Candidate in Human-Computer Interaction
GroupLens Research at the University of Minnesota

Wikimedia Research Showcase
May 20, 2020

@memyselfandHCI

@FauxNeme

New Peer-Reviewed Paper (#CHI2020)

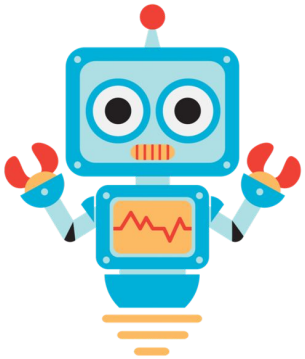
Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems

C. Estelle Smith[1], Bowen Yu[1], Anjali Srivastava[1], Aaron Halfaker[2], Loren Terveen[1], Haiyi Zhu[3]
[1]University of Minnesota, [2]Wikimedia Foundation, [3]Carnegie Mellon University
smit3694@umn.edu, bowen-yu@umn.edu, anjali@umn.edu, ahalfaker@wikimedia.org,
terveen@umn.edu, haiyiz@cs.cmu.edu

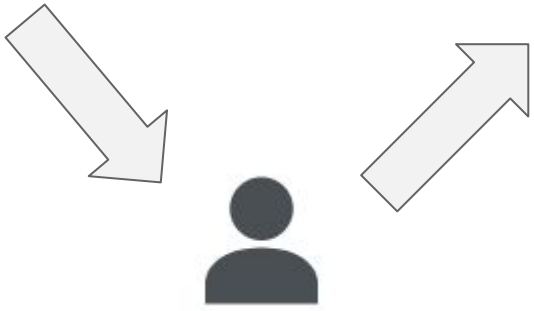
Full open access manuscript:
https://dl.acm.org/doi/abs/10.1145/3313831.3376783

“Human in the loop”

Algorithms to *semi*-automate the task

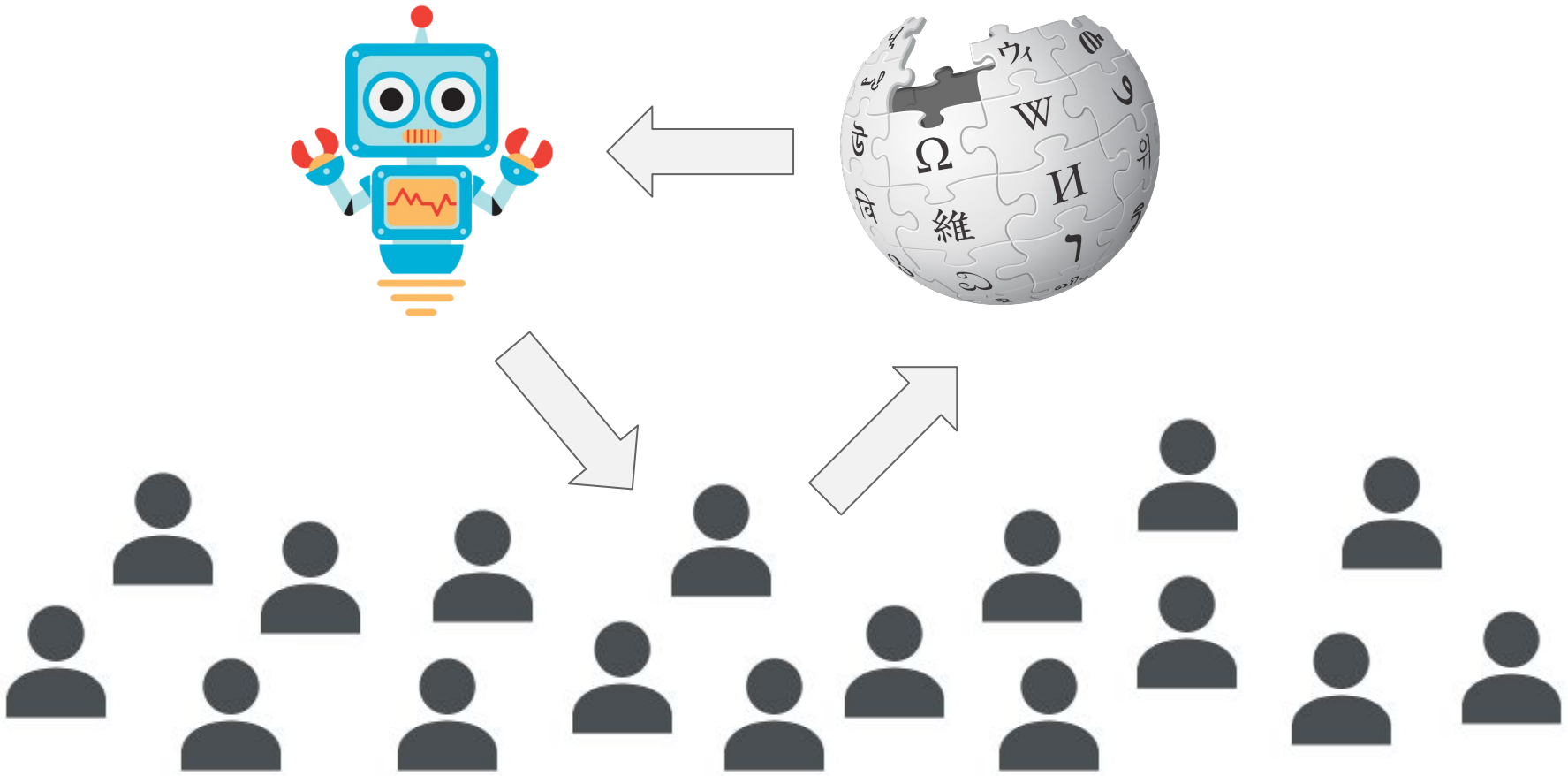


Repetitive Task *(e.g. patrolling Wikipedia for vandalism)*

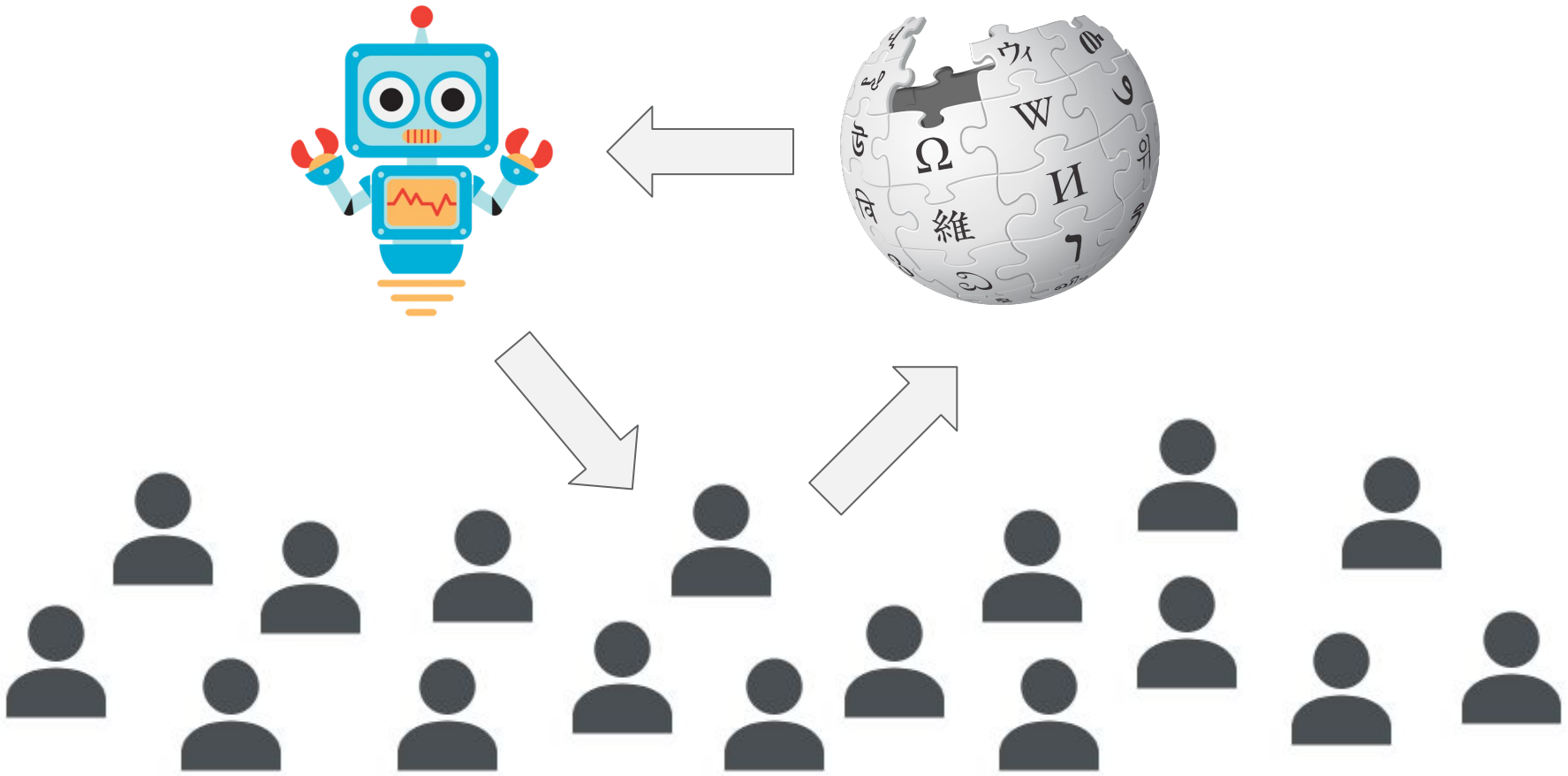


Human judgment is “looped in” to complete the task

Robot-clip-art-book-covers-feJCV3-clipart.png by clipartkid (CC-BY-SA 4.0)



"*Community* in the loop"



How are we currently moderating content generated on Wikipedia?

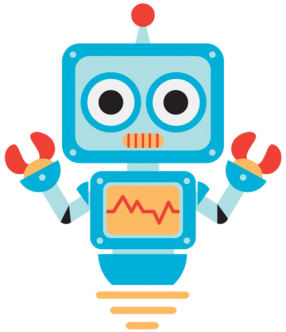
- ML- and non-ML pipelines for reverting damaging edits (Halfaker & Geiger 2012)
- Newcomer motivation hurt by reversion (Halfaker et al. 2011, 2013)
- **ML systems *always* carry risks of unintended consequences**

How to build ML/AI systems for Wikipedia ***without*** *harming the community?*

ORES:
“Objective Revision Evaluation Service”

(Halfaker & Geiger, 2019)

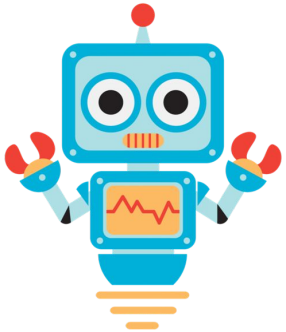
- Online since 2015
- Collection of Machine Learning algorithms
- Web service & API



https://arxiv.org/abs/1909.05189

ORES generates **predictions**

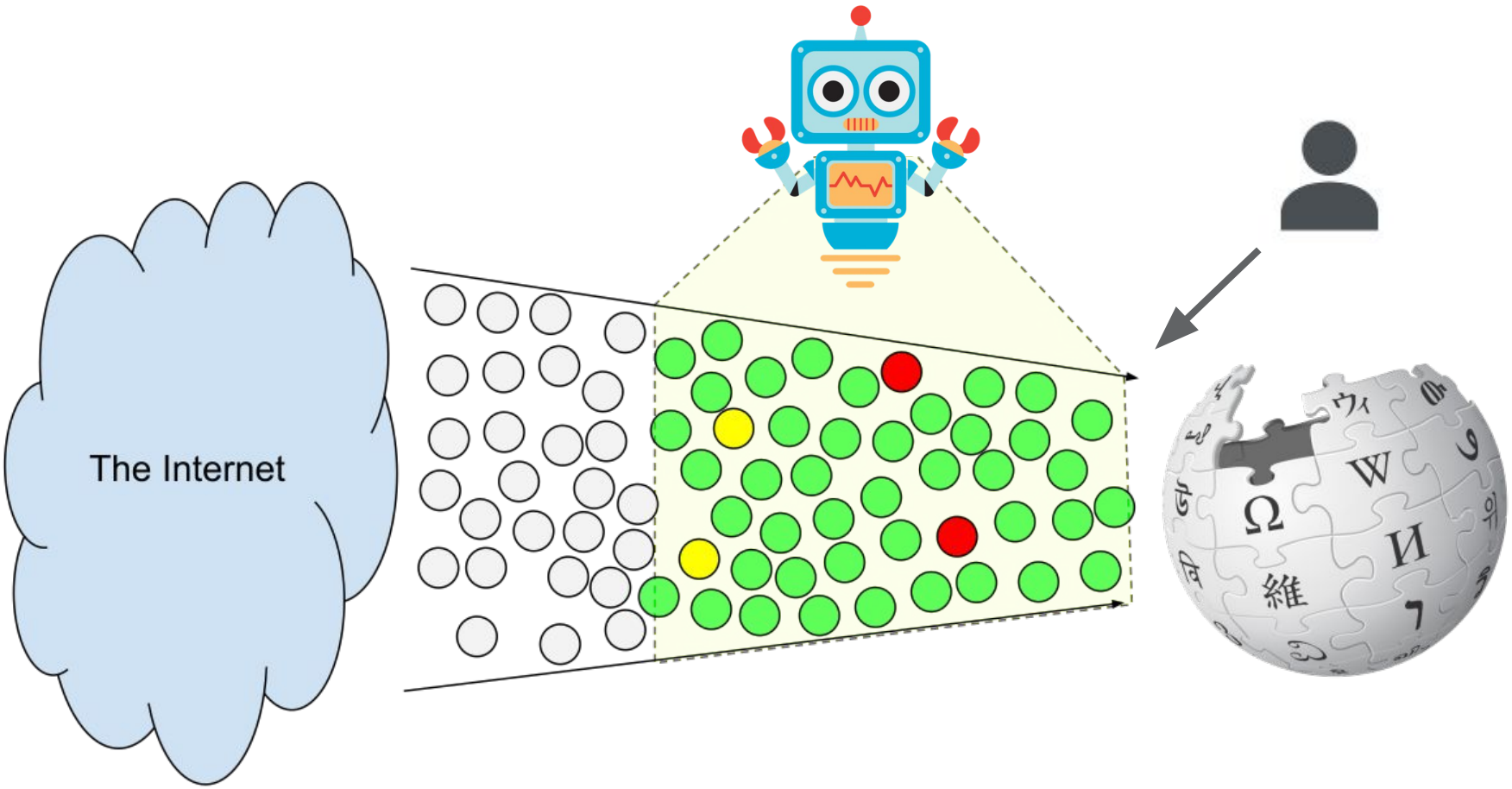
- Edit quality (eg. damaging, goodfaith)
- Draft and article quality
- Draft and article topic



Tools that call ORES

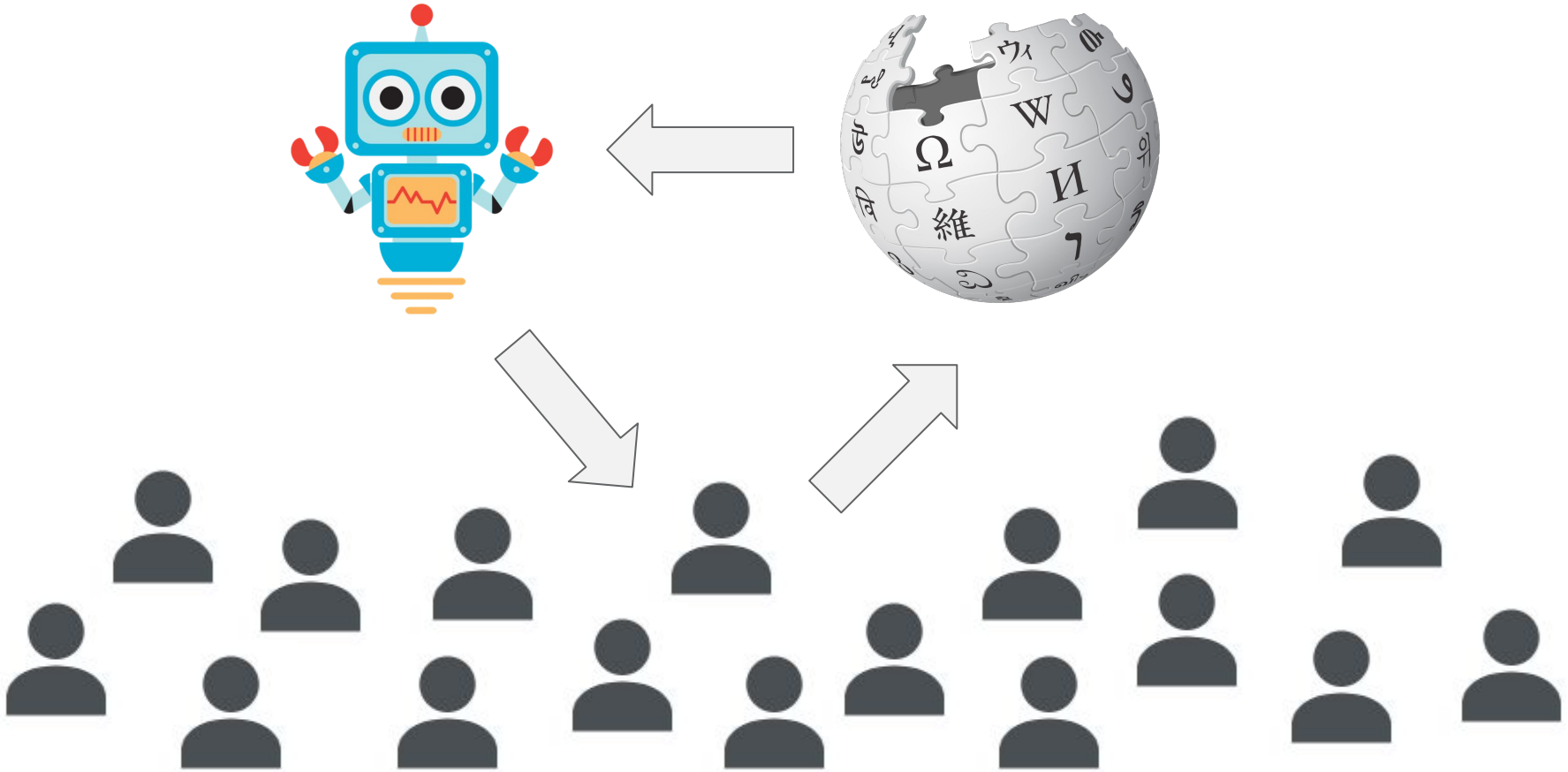
- Recent Changes
- Huggle
- ~30 more here:

  https://www.mediawiki.org/wiki/ORES/Applications



ORES_edit_quality_flow.svg by EpochFail (CC-BY-SA 4.0)

How can we "keep community in the loop" while designing these ORES-based systems?



Value-Sensitive Algorithm Design (VSAD)

(Zhu et al. 2018)

Building from Value-Sensitive Design (VSD),
VSAD is a tripartite approach:

1. Understand **stakeholder values** early in design
2. Use values to **guide algorithm design**
3. Evaluate algorithms on accuracy *and* satisfying **values**

Value-Sensitive Algorithm Design (VSAD)

(Zhu et al. 2018)

Building from Value-Sensitive Design (VSD),
VSAD is a tripartite approach:

1. Understand **stakeholder values** early in design
2. Use values to **guide algorithm design**
3. Evaluate algorithms on accuracy *and* satisfying **values**

We used **qualitative methods** to gather in-depth perspectives across five community stakeholder groups.

Interviews

Participants (16)

- **ORES' Creator** (1)
- **Tool Developers** (2)
- **Wikimedia Product Teams** (4)
- **Editors** (7)
- **Researchers** (2)

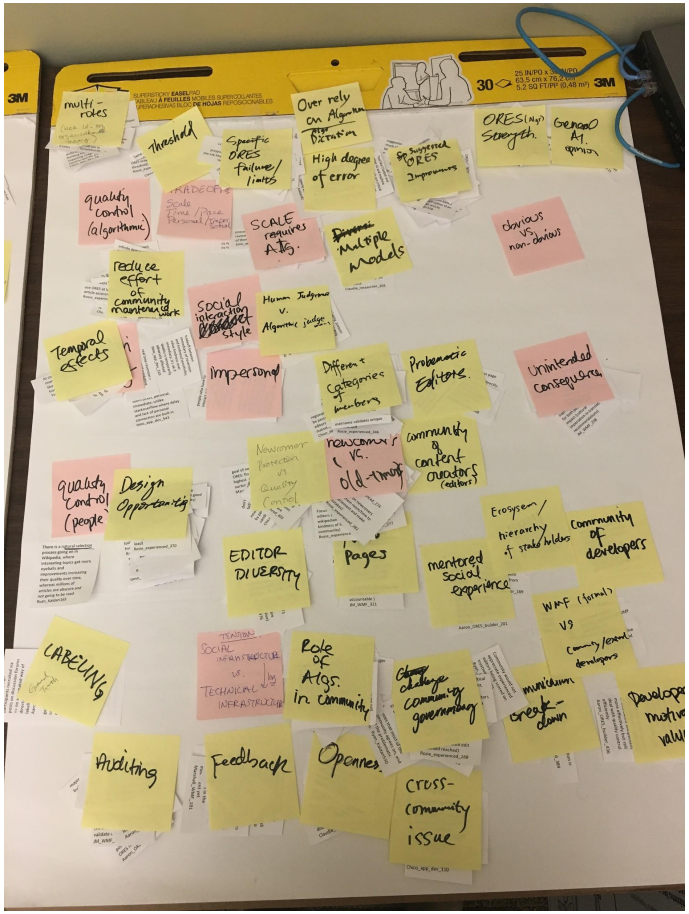
Interviews

Questions

- Role on Wikipedia?
- Experiences related to ORES?
  (Using, building tools, etc.)
- Opinions, ideas for the future?

Analysis using "Grounded Theory Method" (Charmaz 2014)

- Analyze and "code" ***every*** line of interview transcript
- Immersive group meetings to ***cluster*** codes
- ***Discuss and iterate*** on themes



Results

- 2 Creator Values
- 5 Convergent Community Values

- Will discuss a subset of them

Creator Values

1. Enable Consistency and Replicability
2. Facilitate Experimentation

ORES Creator

``What I really hoped to see wasn't that we would do quality control better, exactly, but that more people would start experimenting with quality control tools."

Creator Values

1. Enable Consistency and Replicabi
2. Facilitate Experimentation

ORES Creator

``What I really hoped to see wasn't that we would do quality control better, exactly, but that more people would start experimenting with quality control tools."

``We need to build a contribution platform that allows people to plug their own algorithms in."

WMF

Convergent Community Values

- Data from different stakeholder groups "converged"
- No conflict between stakeholder groups
- However, some intrinsic conflicts between *values themselves*

Convergent Community Values

1. Effort Reduction
2. Human Authority
3. Workflow Support
4. Positive Engagement
5. Community Trust

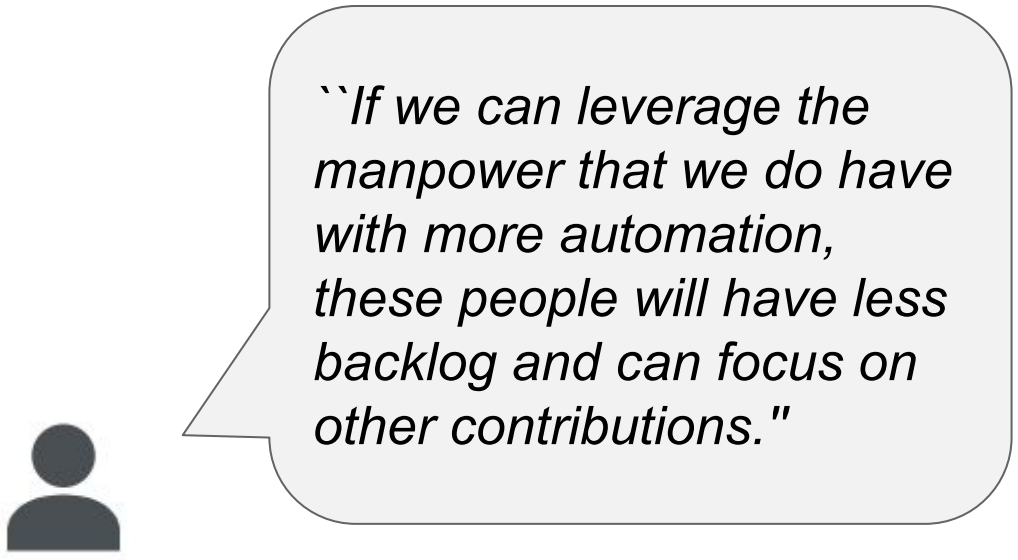
Convergent Community Values

1. Effort Reduction
2. Human Authority
3. Workflow Support
4. Positive Engagement
5. Community Trust

Convergent Community Values

1. Effort Reduction

Reduce the effort of community maintenance.



Convergent Community Values

1. Effort Reduction
2. Human Authority

Maintain Human Judgement as the Final Authority.

Editor

``ORES' purpose is more to create lists of possible problematic pages or edits for human editors to look at, rather than take action fully automatically."

Editor

``I wouldn't rely on ORES 100% of the time. I would still have to use my brain to make a decision."

Convergent Community Values

1. Effort Reduction
2. Human Authority
3. Workflow Support
4. Positive Engagement

Encourage positive engagement with diverse editor groups.

``I think that article quality is driven to a large extent by the diversity of hundreds of users."

Researcher

Convergent Community Values

1. Effort Reduction
2. Human Authority
3. Workflow Support
4. Positive Engagement

Researcher

``I think that article quality is driven to a large extent by the diversity of hundreds of users."

``[The current ecosystem of Wikipedia] limits the diversity of the contributors. So the ecosystem needs to change in order to be more welcoming to certain kinds of people."

WMF

Convergent Community Values

1. Effort Reduction
2. Human Authority
3. Workflow Support
4. Positive Engagement
5. Community Trust

How to *practically* respect these values?

- 25 specific **recommendations** in the paper
- **Consider** values at *every* phase of algorithm development
- Aim to balance **value conflicts**

Value Tensions

Effort Reduction | Positive Engagement

``Because we do everything in real time right now, it's very reactive and very combative. … If [good-faith editors] do something, and a few seconds later, they immediately get a [reversion] notification, maybe those cases can be done later. What's the worst thing that will happen? Maybe somebody will see a syntax error for half an hour, an hour, or maybe even a day. That's maybe not so bad."

Developer

Value Tensions

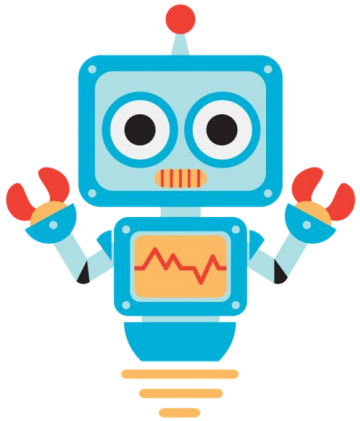
Effort Reduction

Positive Engagement

If interested, BLOG here:
z.umn.edu/wikipediaAI

Three “levels” where we can integrate values

e.g. Human Authority & Effort Reduction

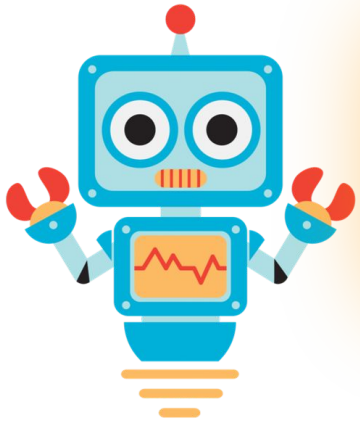


Algorithm

User Interface

Work Process

Three “levels” where we can integrate values



Algorithm

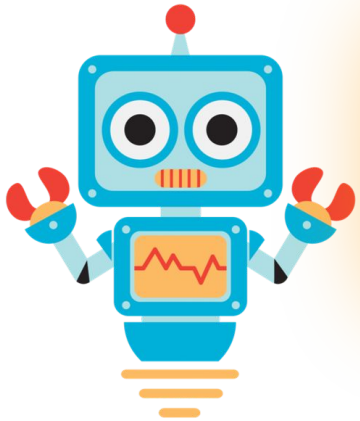
“Human Authority” vs.
“Effort Reduction”

...to tweak algorithmic parameters.
(precision, recall, etc.),

Three “levels” where we can integrate values

min false-negative

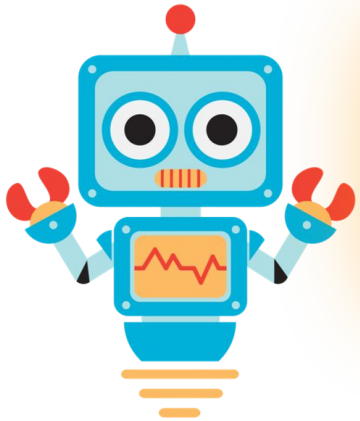
Counter vandalism bot



Algorithm

positive == damaging edit
negative == good quality edit

Three “levels” where we can integrate values



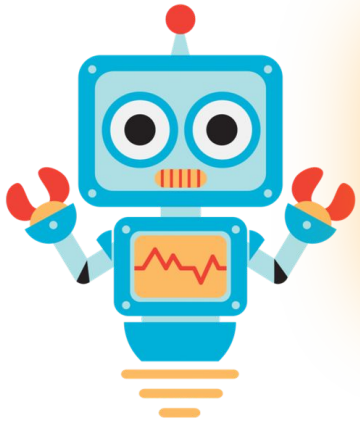
Algorithm

min false-negative
Counter vandalism bot

min false-positive
Semi-automated edit review

positive == damaging edit
negative == good quality edit

Three “levels” where we can integrate values



Algorithm

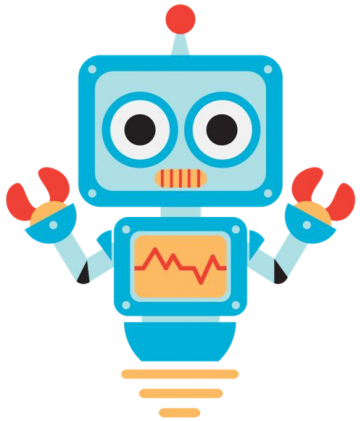
positive == damaging edit
negative == good quality edit

min false-negative
Counter vandalism bot

min false-positive
Semi-automated edit review

min false-negative
-for-newcomers
& overall-error < 0.1
Newcomer protection in quality control

Integrating Values in Algorithmic System Design



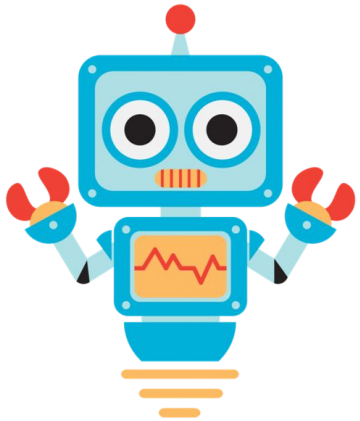
Algorithm

User Interface

Interactive Visualization

- Inputs → Outputs
- Predictions & Errors
- Better tools

Three “levels” where we can integrate values

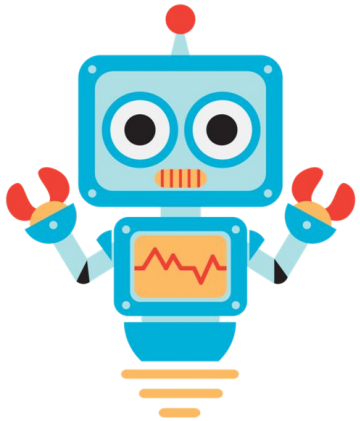


Algorithm

User Interface

Work Process

Three “levels” where we can integrate values



1. Automated bot reverts highly likely damaging edits
2. Semi-automated review / reversion
3. Socialization tool

Work Process

Value Tensions

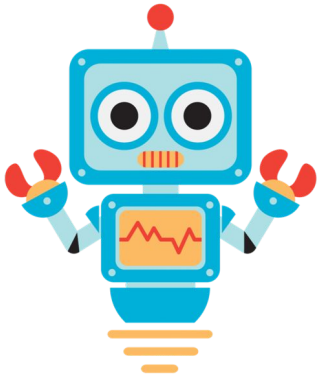
WIKIMEDIA FOUNDATION

Facilitate Experimentation

Effort Reduction
Human Authority
Workflow Support
Positive Engagement
Community Trust

``We definitely don't want to put barriers between people using ORES, [but] if you can use ORES, you can also use ORES inappropriately."

ORES Creator

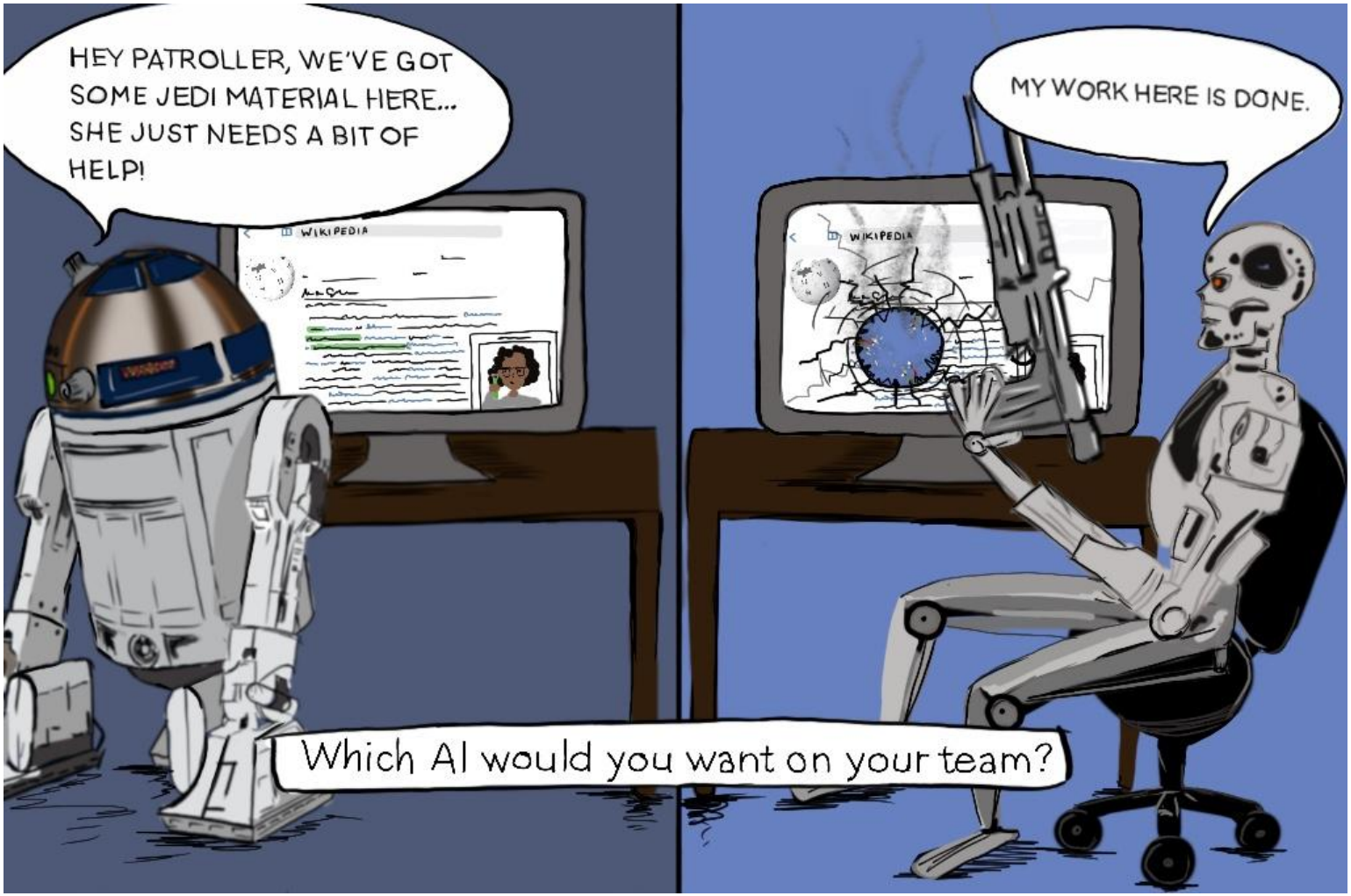
Questions?  Comments?

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*

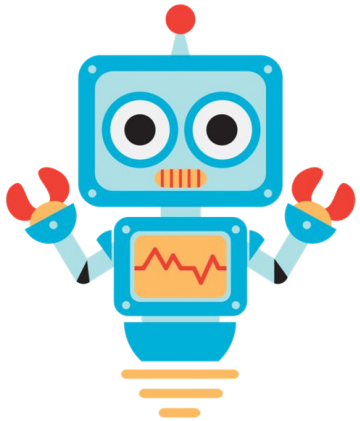
C. Estelle Smith
smit3694@umn.edu
@memyselfandHCI

[Q&A SLIDES]

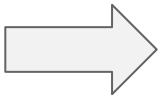


z.umn.edu/wikipediaAI

Three “levels” where we can integrate values



1. Automated bot reverts highly likely damaging edits
2. Semi-automated review / reversion
3. Socialization tool



1. Automated bot reverts highly likely damaging edits
2. Semi-automated review (no reversion)
3. Socialization tool
4. (Possible reversion)

OR, also considering Positive Engagement...

Work Process

Convergent Community Values

1. Effort Reduction
2. Human Authority
3. Workflow Support
4. Positive Engagement
5. Community Trust

Support differing peoples' differing workflows.

Establish the trustworthiness of people and algorithms within the community.

Value-Sensitive Algorithm Design (Zhu et al. 2018)

A value is...

``What a person or group of people consider important in life."

(Borning & Muller 2012)

Convergent Community Values

1. Effort Reduction
2. Human Authority
3. Workflow Support
4. Positive Engagement

"Evolve the ecosystem"

- Understand when actions taken by humans vs. AIs
- Algorithms should *facilitate,* not replace, socialization
- Transparent explanations of algorithms
- Algorithms should help share ways to *grow* in the community

Workflow Support

Algorithmic tools should facilitate workflows that help to achieve users' actual end goals.

Developers should identify sets of users' priorities throughout their workflows, and build tools that are configurable to those different priorities.

Developers should create intuitive UI/UX elements that make it easy to select workflows based on users' different priorities.

UI/UX elements in algorithmic tools should be designed to give users the flexibility to select and stay focused on the type of use case they want to work on, until they decide to switch to a different one.

Positive Engagement

Users should be able to understand which actions were taken by algorithms, which actions were taken by humans, and how to contest decisions.

Social connections within the community should be facilitated rather than replaced or weakened by algorithmic systems.

Algorithmic systems should provide transparent explanations of their behavior, and accessible training resources for effective interactions with them.

Algorithmic systems should provide and recommend helpful ways for users to learn and grow within the community.

Community Trust

Developers should continuously engage with the communities affected by algorithmic systems to build and maintain trust.

To aid in community governance efforts, algorithmic systems should provide mechanisms to assess the trustworthiness of community members based on their community contributions and behaviors.

Trusted users should be able to impact algorithms by providing feedback on their performance, even if they don't understand all details of how the algorithms work.