Community-Centered AI Design: Incorporating Community Values into the Development of AI Technologies

Haiyi Zhu

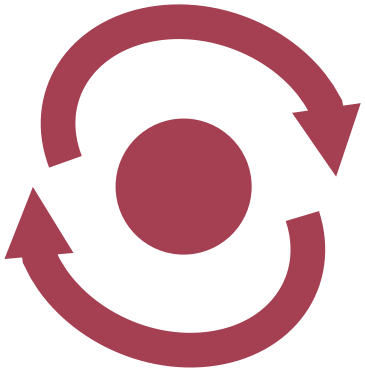
Assistant Professor

Human Computer Interaction Institute

Carnegie Mellon University

My Research

AI Technologies



Groups Communities Organizations

AI and Communities

- AI-Supported Content Moderation

Community context: Wikipedia

- Peer Support Mental Health Community

Community context: 7 Cups

- AI-Supported Child Protection

Community context: Allegheny County in Pennsylvania

- Empowering and Enhancing Gig Workers

Community context: Uber, Lyft, Upwork, Freelancer.com, etc

Research Focus

Community-centered AI Design:

- How can we better incorporate communities' values and goals into the design of AI systems?

*NSF CHS Core Program
*NSF EAGER on AI and Society
*NSF Fairness in AI
*NSF Smart and Connected Communities

Today's talk

- AI-Supported Content Moderation

Community context: Wikipedia

- Peer Support Mental Health Community

Community context: 7 Cups

- AI-Supported Child Protection

Community context: Allegheny County in Pennsylvania

- Empowering and Enhancing Gig Workers

Community context: Uber, Lyft, Upwork, Freelancer.com, etc

Background

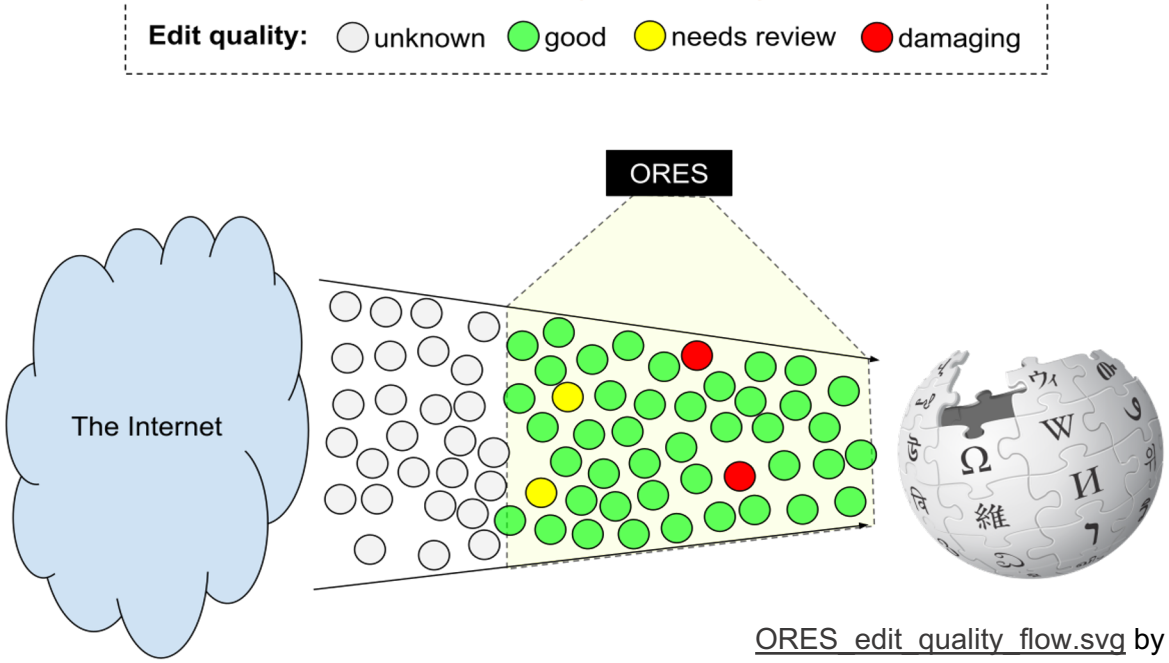
Wikipedia

In Wikipedia, 160,000 edits are made per day by various groups of editors. Quality control is one of the major tasks faced by the wikipedia community.

Wikipedia Logo 1.0 by by Nohat (CC-BY-SA 4.0)

Objective Revision Evaluation System (ORES) generates **predictions**

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

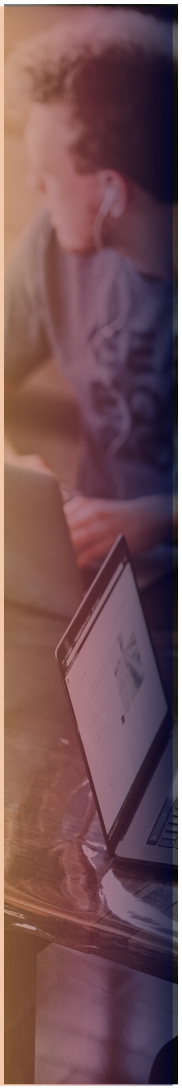


ORES_edit_quality_flow.svg by EpochFail (CC-BY-SA 4.0)

Tools that call ORES

- Recent Changes
- Huggle
- ~30 more here:

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



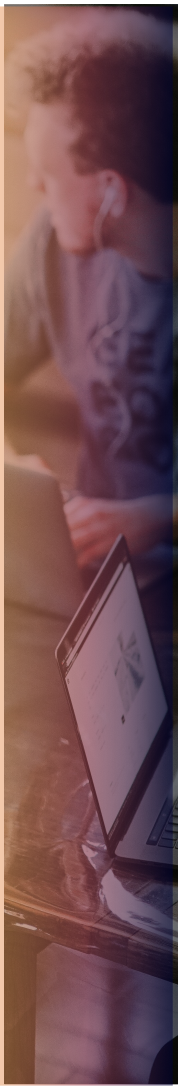
Today's Talk

1. Understanding Wikipedia community values for quality prediction systems (Smith et al, CHI 2020)

2. Creating visualization systems to capture and explain the trade-offs between multiple community values (Ye et al, DIS 2021)

3. Conducting community workshops that allow community stakeholders to discuss and negotiate the trade-offs (Ongoing work)

4. Big Picture



Today's Talk

1. Understanding Wikipedia community values for quality prediction systems (Smith et al, CHI 2020)

2. Creating visualization systems to capture and explain the trade-offs between multiple community values (Ye et al, DIS 2021)

3. Conducting community workshops that allow community stakeholders to discuss and negotiate the trade-offs (Ongoing work)

4. Big Picture

Interviews

Participants (16)

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

Interviews

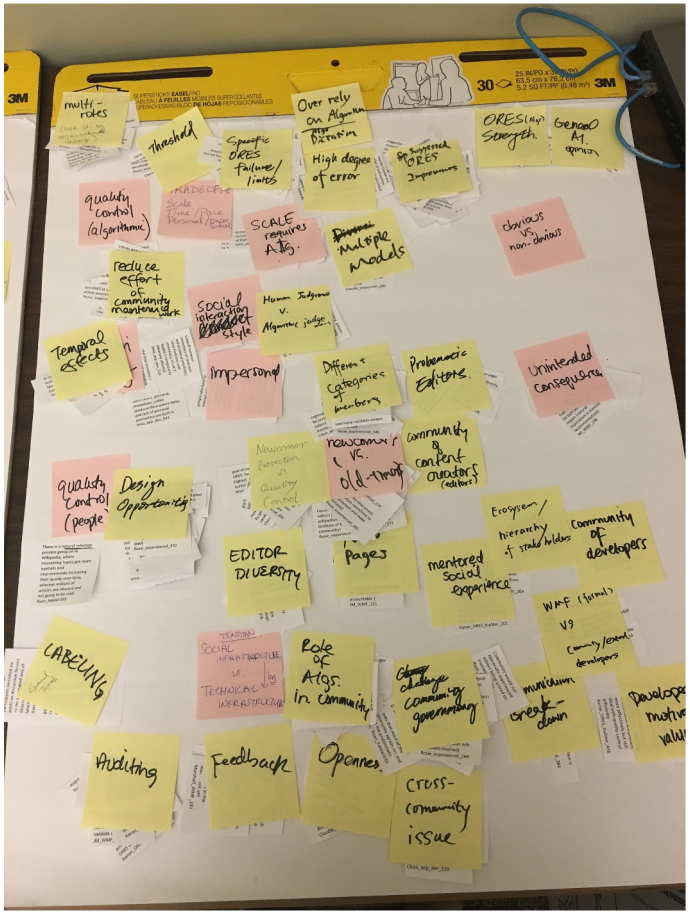
speaking.svg by MScharwies (CC BY-SA 4.0)

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

Convergent Community Values

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

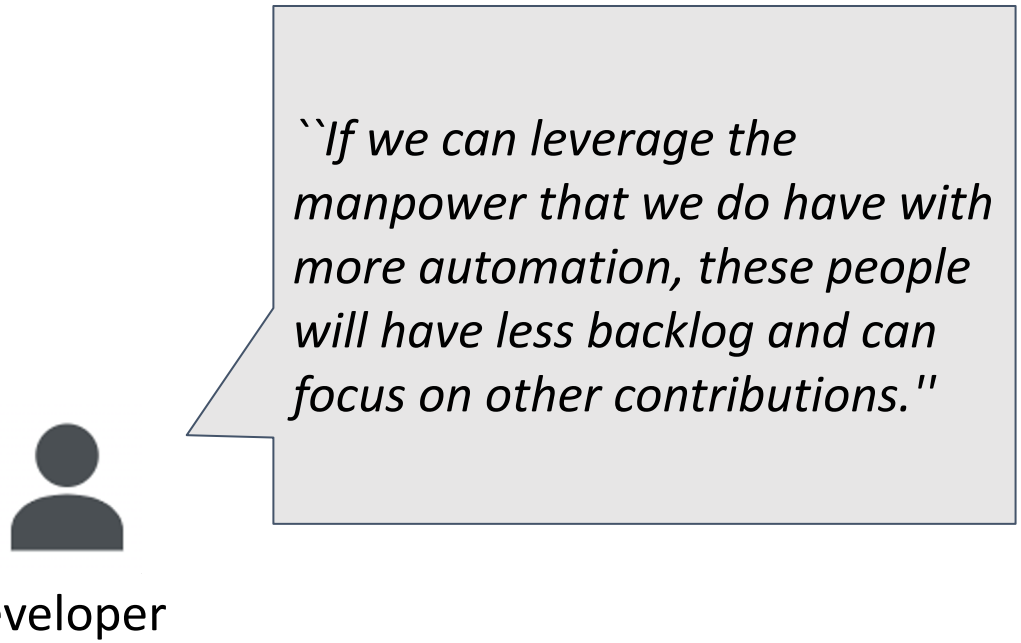
Convergent Community Values

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
3. Workflow Support
4. Positi

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. Positi

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

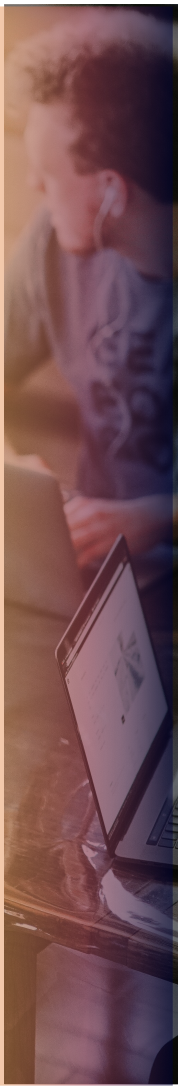
Value Tensions

Effort Reduction

1. High overall accuracy
2. Low false negatives (i.e., catching all the possible damaging edits)

Positive Engagement with diverse editor groups

1. Low false positives (e.g., not falsely labeling good edits)
2. Low disparity between model performance on different editor groups



Today's Talk

1 Understanding Wikipedia community values for quality prediction systems (Smith et al, CHI 2020)

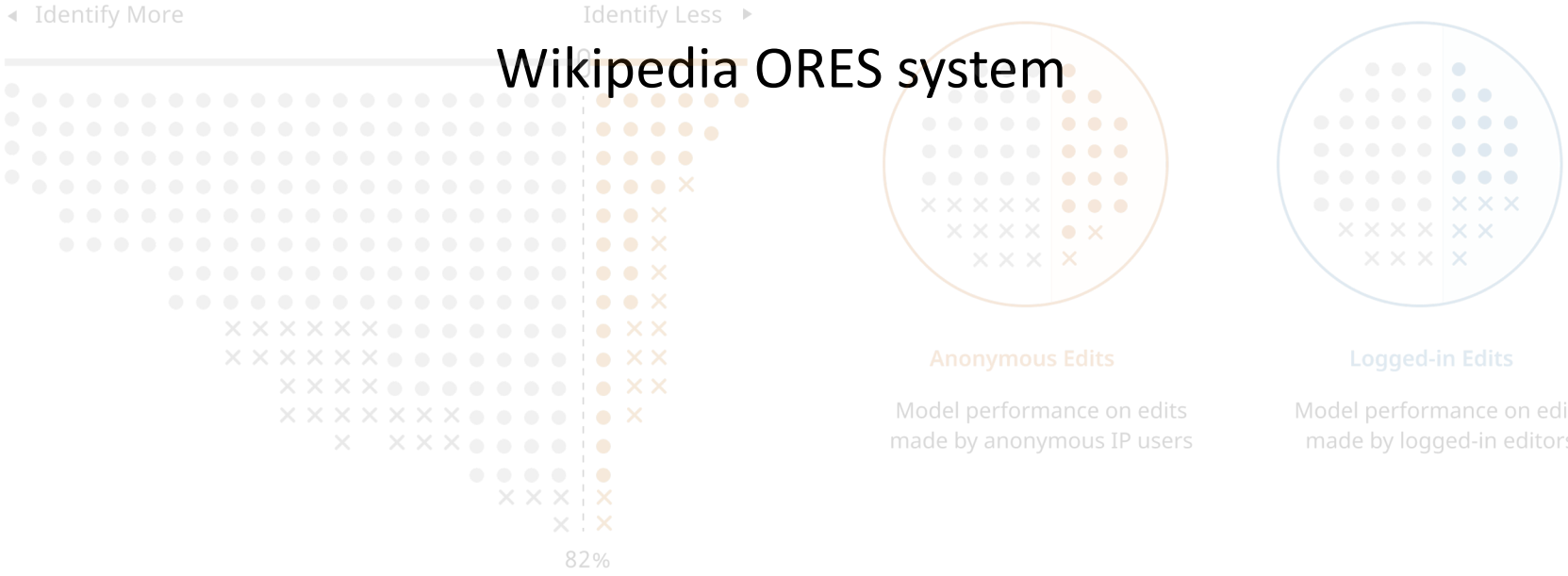
2 Creating visualization systems to capture and explain the trade-offs between multiple community values (Ye et al, DIS 2021)

3 Conducting community workshops that allow community stakeholders to discuss and negotiate the trade-offs (Ongoing work)

4 Big Picture

Goal

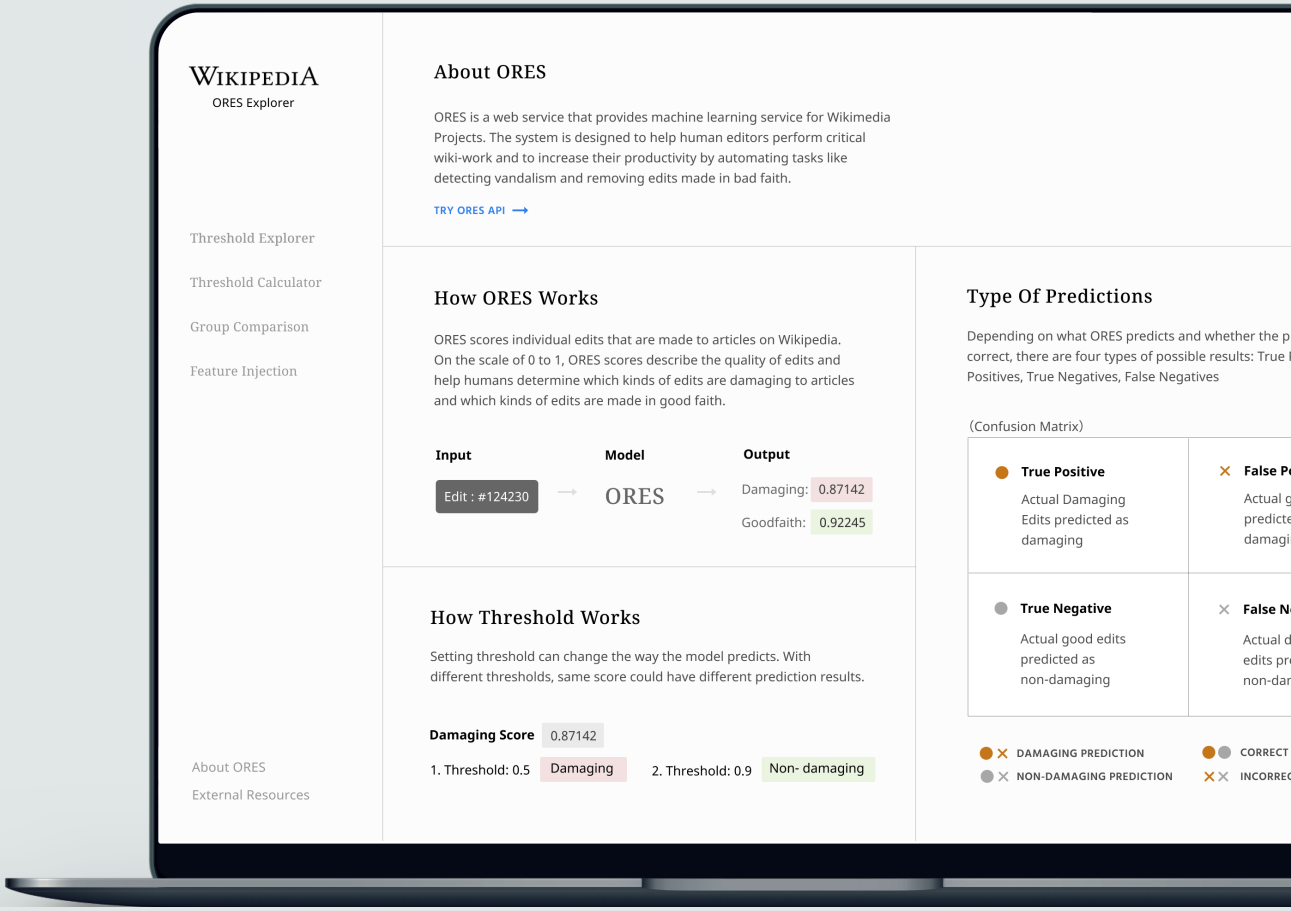
ORES Explorer: A set of visualizations to help application designers and community members to understand the inherent tradeoffs in Wikipedia ORES system



Iterative Design Process

About ORES

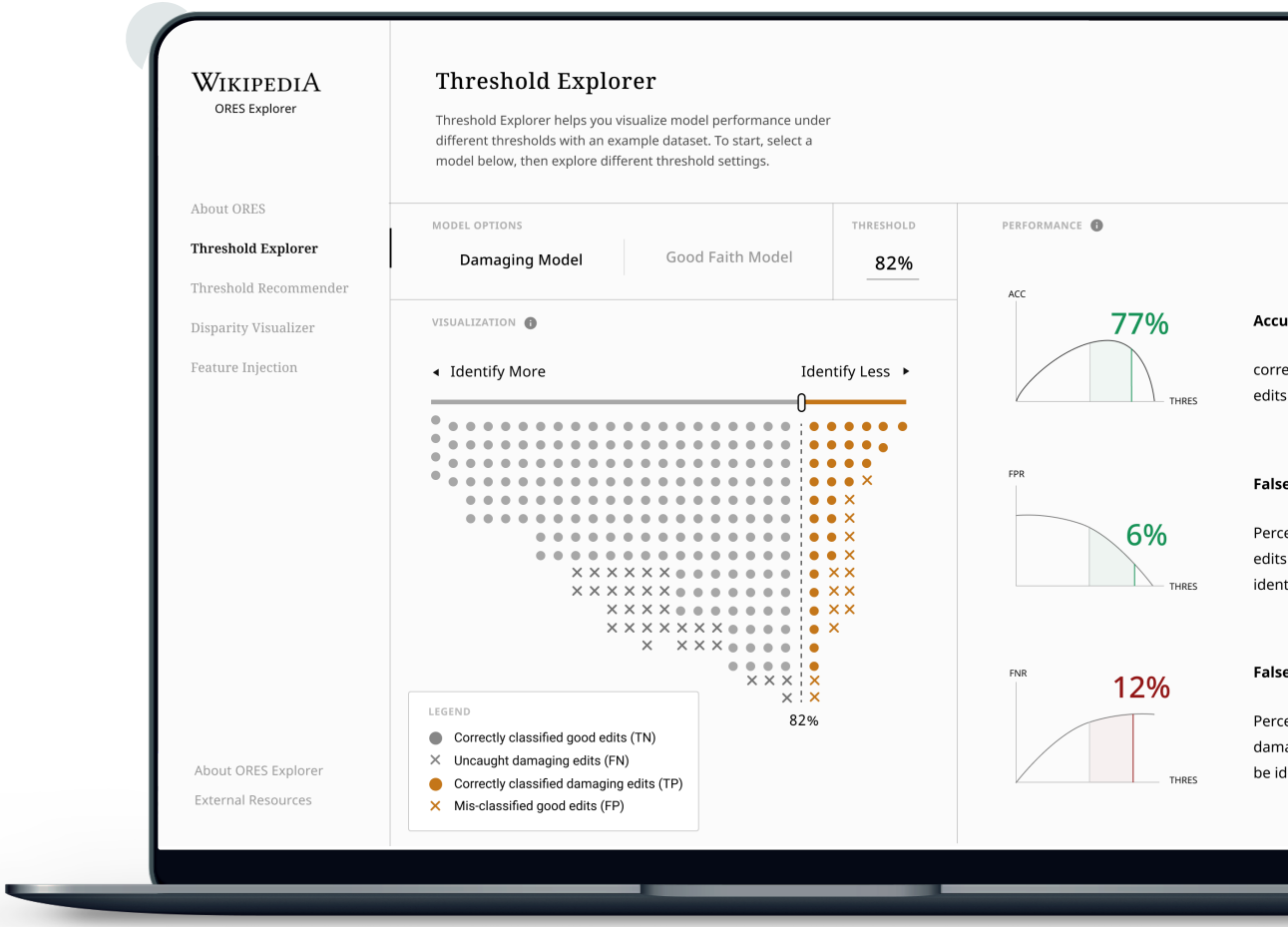
Provide basic background information on how ORES works with the necessary ML knowledge.



Threshold Viz

Threshold Explorer

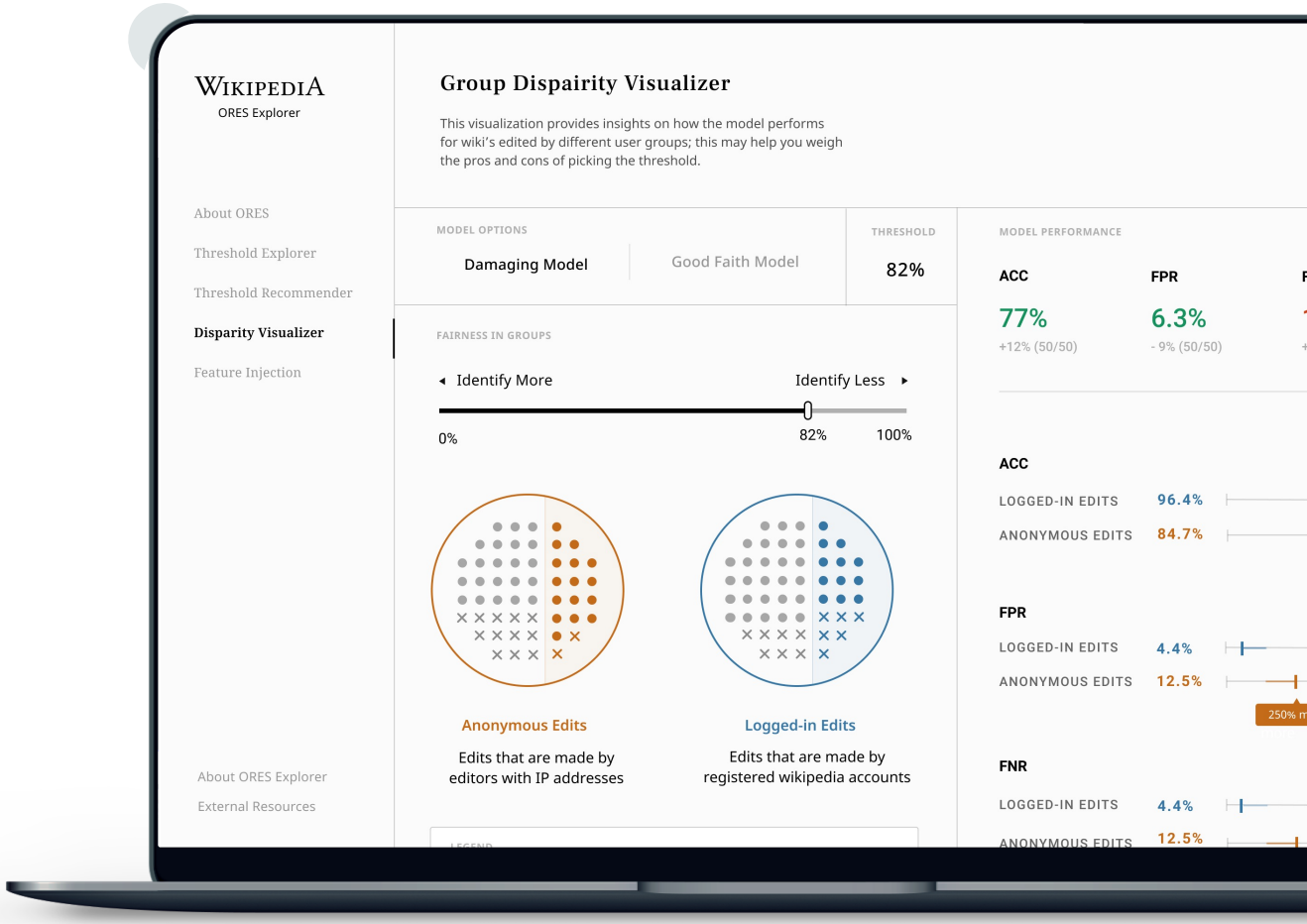
Visualize the impact and trade-offs of setting different thresholds



Fairness Viz

Group Disparity Visualizer

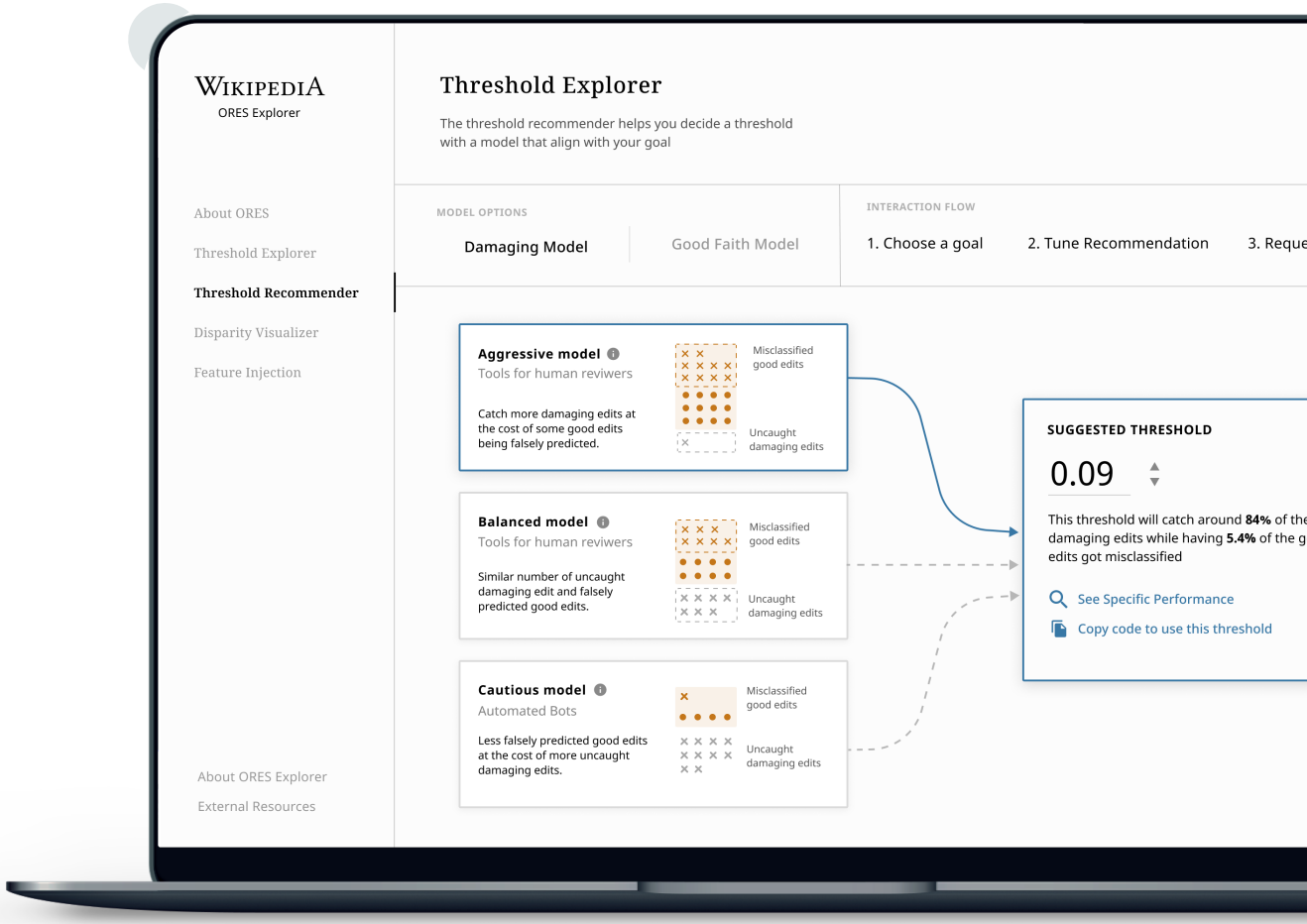
Visualize ORES's performance on different groups of editors



Threshold Viz

Threshold Recommender

Recommend threshold based on users' goals and preferences



Haiyi Zhu, Research Overview

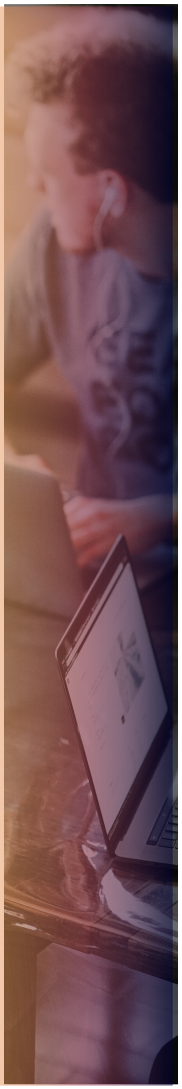
Participants Recruitment & User Test Protocol

ID	Group	Role
P1	Outside Wikipedia	Associate UX Designer, Spotify
P2	Outside Wikipedia	Product Designer, Facebook
P3	Outside Wikipedia	Digital Product Manager, CitiGroup
P4	Outside Wikipedia	Student, Harvard Graduate School of Design
P5	Outside Wikipedia	UX Designer, Amazon
W1	Wikipedia Member	Edit Reviewer, Recent Changes Tool User
W2	Wikipedia Member	Volunteer Editor, Implemented ORES at eu.wikipedia.org
W3	Wikipedia Member	Developer, Researcher, Volunteer for Wikimedia Projects
W4	Wikipedia Member	Researcher, Edit Reviewer, Recent Changes Tool User
W5	Wikipedia Member	Huggle Application Designer

Table 1: Participant Information.

Evaluation

- ORES Explorer improved participants' understanding of the trade-offs in setting different ORES model thresholds and the associated impacts.
- Although Group Disparity Visualizer helped surface the ORES model's performance disparity in different editor groups, most participants accepted the disparity as a natural occurrence and were not concerned about fairness implications in the system.



Today's Talk

1 Understanding Wikipedia community values for quality prediction systems (Smith et al, CHI 2020)

2 Creating visualization systems to capture and explain the trade-offs between multiple community goals (Ye et al, DIS 2021)

3 Conducting community workshops that allow community stakeholders to discuss and negotiate the trade-offs (Ongoing work)

4 Big Picture

Goal

Combining visualization and community deliberation workshops to explain the tensions to the community members and resolve the tensions between the community goals in the ORES systems.

- **Dutch** Community
- **English** Community

Deliberation Protocol

1. Complete a **presurvey** (10 minutes)
2. Explore the **ORES explorer interface** and create **1** model card (20 minutes)
3. **Group Discussion** (40 minutes) and proposal **writing** (10 minutes)
4. Complete a **postsurvey** (10 minutes)

WIKIPEDIA

ORES Explorer

LeijieWang

About ORES

Explore the Model

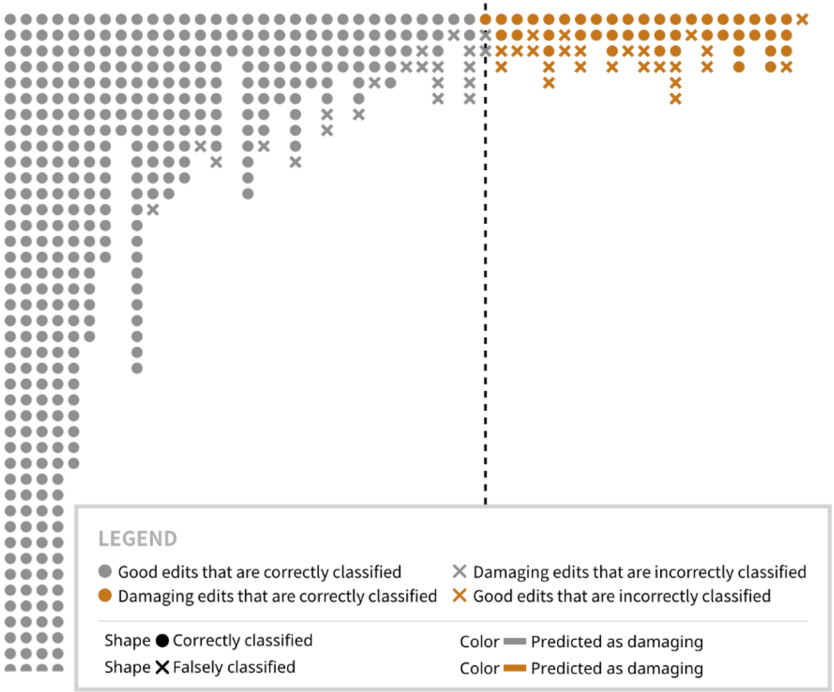
Choose Your Modelcard

Your Modelcard History

OVERALL DISPARITY

Setting threshold can change the way the damaging model predicts. Only scores higher than the threshold will be predicted as damaging. Move the slide to how ORES performs for each threshold and click on a revision (cross or circle) to view more details.

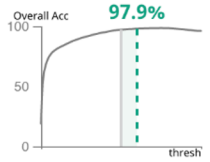
◀ Identify More as Damaging 60% Identify Less as Damaging ▶



Accuracy

How many edits are correctly classified

Overall 97.9%

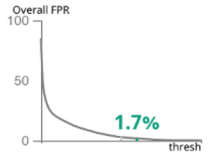


Newcomer 95.7%
Experienced 99.8%
Anonymous 91.3%

False Positive Rate

How many good edits are incorrectly classified

Overall 1.7%

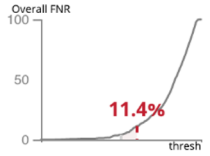


Newcomer 3.9%
Experienced 0.0%
Anonymous 8.7%

False Negative Rate

How many damaging edits are incorrectly classified

Overall 11.4%

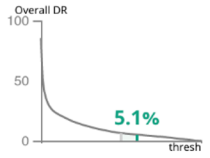


Newcomer 7.0%
Experienced 41.8%
Anonymous 9.0%

Damaging Rate

How many edits are classified as damaging

Overall 5.1%



Newcomer 14.4%
Experienced 0.3%
Anonymous 20.0%

Haiyi Zhu, Research Overview

WIKIPEDIA

ORES Explorer

LeijieWang

About ORES

Explore the Model

Choose Your Modelcard

Your Modelcard History

Choose Your Model Card

This section helps you examine the model performance of a given threshold in a model card, and it also provides you a place to write down your rationale for your choice of models.

FAIRNESS IN GROUPS

◀ Identify More as Damaging — Iden 72% Less as Damaging ▶

Accuracy	False Positive Rate
Edits that are correctly classified	Good edits that are incorrectly classified
Overall: 98.3%	Overall: 0.8%
Newcomer: 95.9%	Newcomer: 1.9%
Experienced: 99.7%	Experienced: 0.0%
Anonymous: 93.9%	Anonymous: 3.9%
False Negative Rate	**Damaging Rate**
Damaging edits that are incorrectly classified	Edits that are classified as damaging
Overall: 23.8%	Overall: 3.7%
Newcomer: 20.5%	Newcomer: 11.1%
Experienced: 61.2%	Experienced: 0.2%
Anonymous: 19.9%	Anonymous: 14.4%

EXPLANATION

Why Do You Choose This Model?

Write down your rationale and then click the "Save" button. Then you could share your model cards with others simply by the generated link or wikitext!

SAVE

Haiyi Zhu, Research Overview

WIKIPEDIA

ORES Explorer

Performance

THE PERFORMANCE OF THE CHOSEN MODEL

THRESHOLD 72% BY LeijieWang

Accuracy

Edits that are correctly classified

Overall: 98.3%

Newcomer: 95.9%

Experienced: 99.7%

Anonymous: 93.9%

False Positive Rate

Good edits that are incorrectly classified

Overall: 0.8%

Newcomer: 1.9%

Experienced: 0.0%

Anonymous: 3.9%

False Negative Rate

Damaging edits that are incorrectly classified

Overall: 23.8%

Newcomer: 20.5%

Experienced: 61.2%

Anonymous: 19.9%

Damaging Rate

Edits that are classified as damaging

Overall: 3.7%

Newcomer: 11.1%

Experienced: 0.2%

Anonymous: 14.4%

Explanation

THE REASON WHY THIS MODEL WAS CHOSEN BY LEIJIEWANG

I like this model because it has a relatively lower false positive rate and the accuracy is not so bad for each group.

SHARE CHOOSE ANOTHER

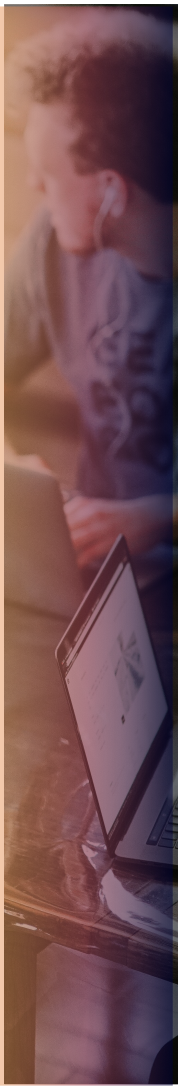
Group Discussion: Which model you think is producing the best outcomes and would recommend for the community to use.

What is your definition of good outcomes in Wikipedia?
What are the pros and cons of different models?
If you're going to develop a model for the English/Dutch Wiki, which one would your group collectively choose?

Writing a group proposal about your collectively chosen model.

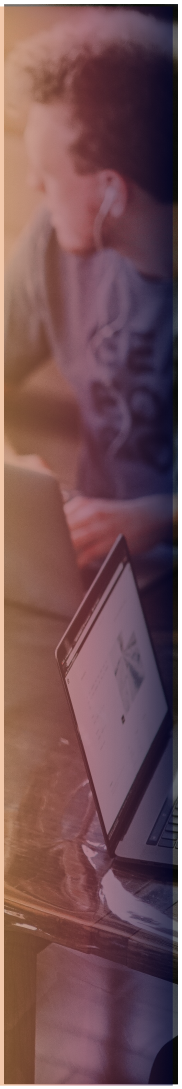
If you agree on a model, please write down your rationale. If you are not able to agree upon a model, please also share your reasons.

What are the high-level principles (values) that the ORES developers (or any future AI system builders at Wiki) should consider to better benefit the English/Dutch Wikipedia?



Summary

1. Contribute to a broader understanding of **human values related to AI-supported governance** in online communities.
2. Design novel approach (visualizations and deliberation workshops) that facilitates greater **community control** and **community agency** in AI design.
3. The work has direct implications on the improvement of Wikipedia's ORES, as well as broader implications on the design of content moderation systems for other communities such as Reddit and Twitch.



Today's Talk

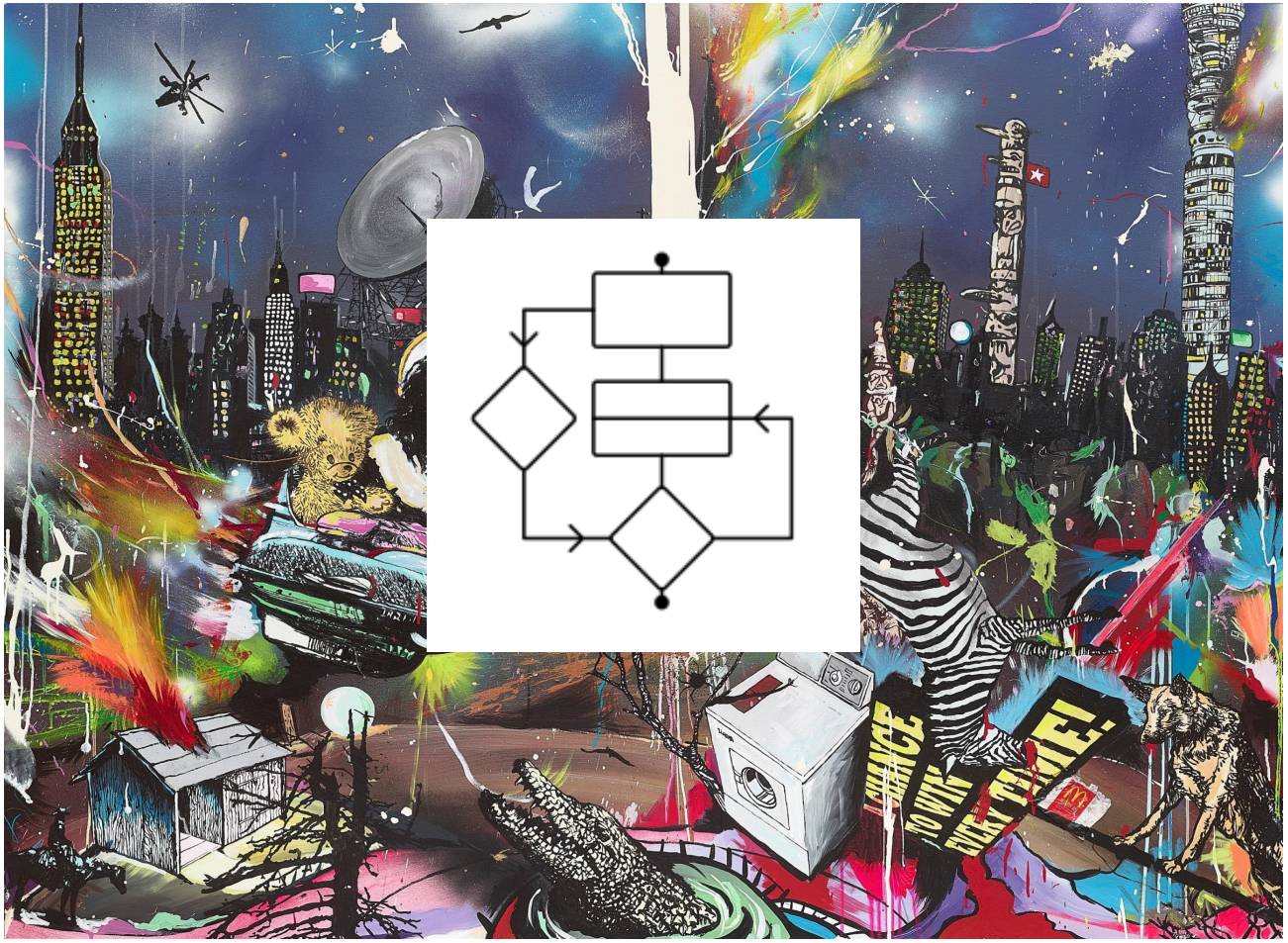
1 Understanding Wikipedia community values for quality prediction systems (Smith et al, CHI 2020)

2 Creating visualization systems to capture and explain the trade-offs between multiple community goals (Ye et al, DIS 2021)

3 Conducting community workshops that allow community stakeholders to discuss and negotiate the trade-offs (Ongoing work)

➡ **4 Big Picture**

Big Picture: Apply AI to Address Communities' Problems



Dan Baldwin's Faith Less

https://www.oxfordtimes.co.uk/news/14302799.vivid-images-of-a-messy-world-from-artist-dan-baldwin-at-cassington-gallery/

Goals

Work closely with multiple online and offline communities and solve their problems

Reflect on and position our insights

Framework for community-centered AI Design

Generate a set of best practices, methods and metrics, and create and publish a set of resources as components of a framework to support replication in community-centered AI work

Acknowledgements

Participants in the studies!

Aaron Halfaker; Ronald and Ciell (Dutch Wikipedia)

Estelle Smith, Ethan Ye, Hong Shen, Anna Yuan, Shaurya Gaur, Bowen Yu, Leijie Wang, Steven Wu, Mark Snyder, Jodi Forlizzi, Raghav Karumur, Loren Terveen

Grouplens Lab@UMN and HCII@CMU

Papers

C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). ACM, New York, NY, USA, 1-14. (Link)

Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-offs Across Multiple Objectives. In Proceedings of the 2020 on Designing Interactive Systems Conference (DIS'20). ACM, New York, NY, USA. (Link)

Zining Ethan Ye, Xinran Yuan, Shaurya Gaur, Aaron Halfaker, Jodi Forlizzi, and Haiyi Zhu. 2021. Wikipedia ORES Explorer: Visualizing Trade-offs For Designing Applications With Machine Learning API. In Designing Interactive Systems Conference 2021 (**DIS '21**), June 28-July 2, 2021, Virtual Event, USA. ACM, New York, NY, USA, 12 pages.

Questions!

Haiyi Zhu

WWW.HAIYIZHU.COM

haiyiz@cs.cmu.edu