Towards Bridging and Governing Decentralized Communities

Belén Saldías

belen@mit.edu, belencarolina.com

PhD Defense

Media Arts and Sciences

@ Massachusetts Institute of Technology

January 22, 2025

Dissertation committee

Deb Roy, Ph.D.

Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Rosalind Picard, Sc.D.

Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Jonathan Zittrain, J.D.

George Bemis Professor of International Law

Harvard University

Social media can bridge us

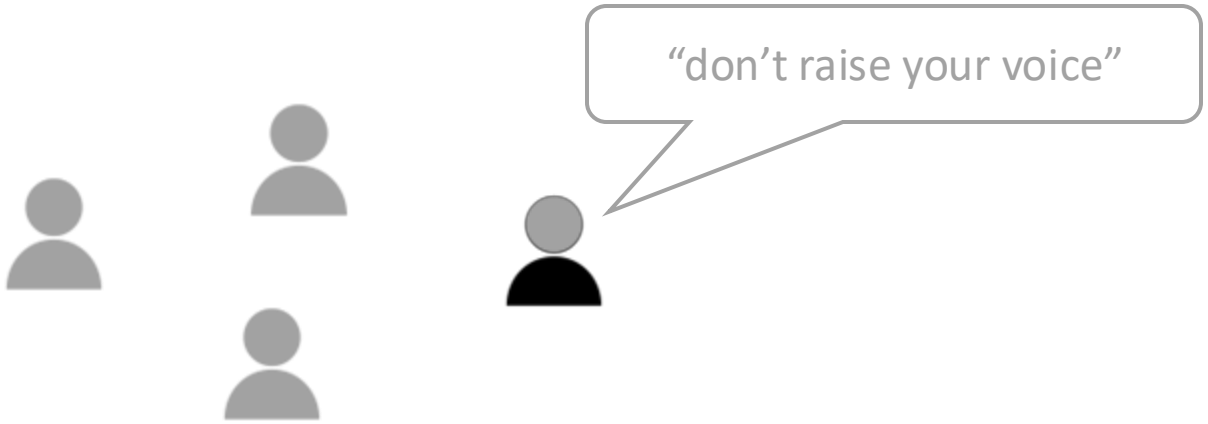
Social media can bridge us

but rather it’s dividing us.

Social media can bridge us

but rather it’s dividing us.

How can we be more intentional in designing online spaces?



Family

Work

Music band

Sports team

Family

“you must salute others”

Work

Music band

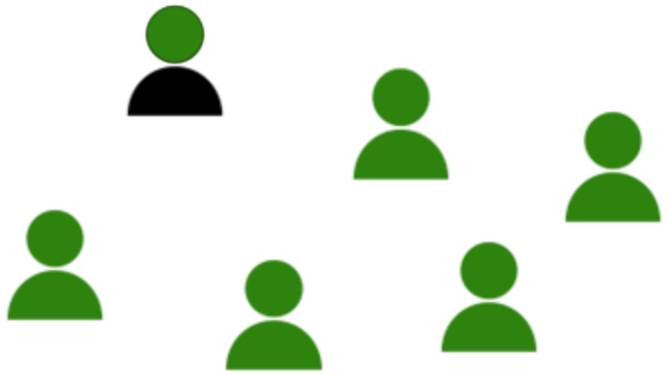
Sports team

Family

Work

“make sure you yell my name”

Music band



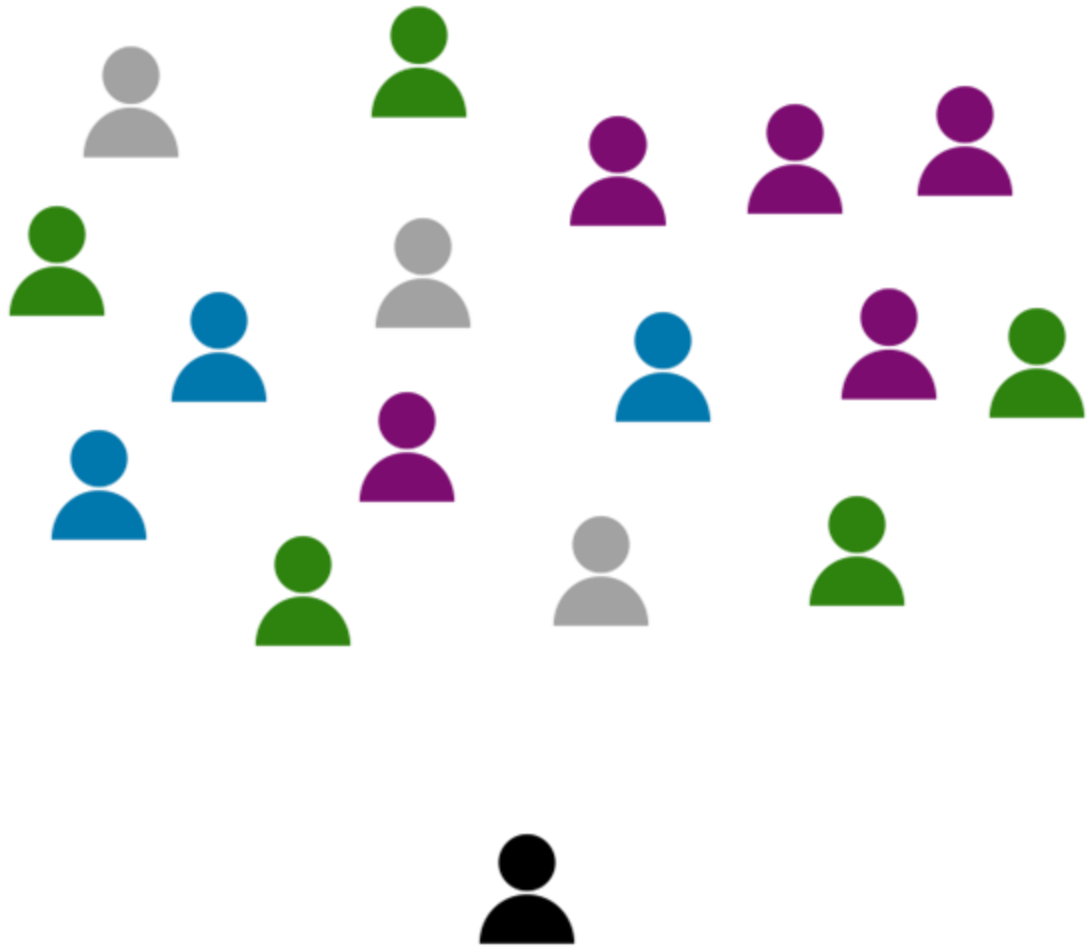
Sports team

Family

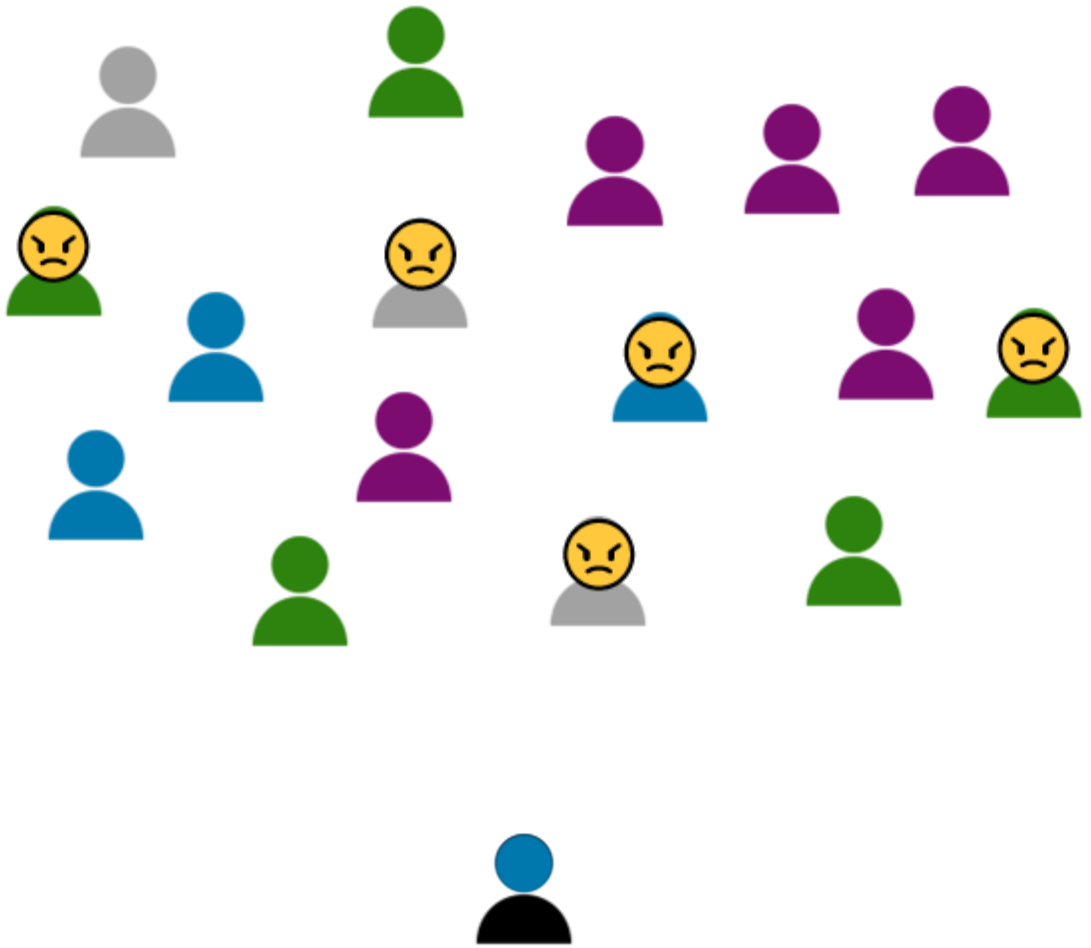
Work

Music band

Sports team



Context collapse

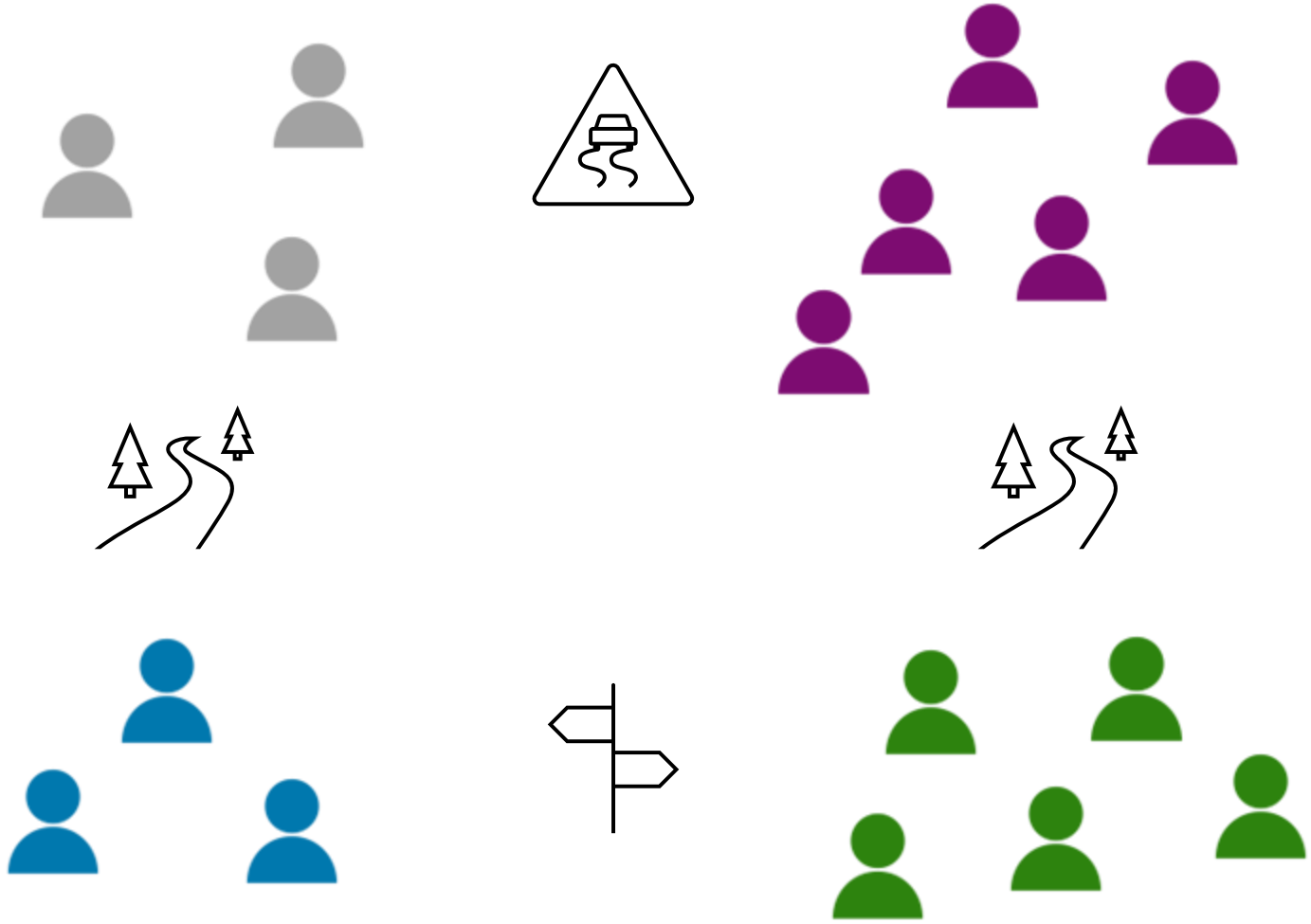


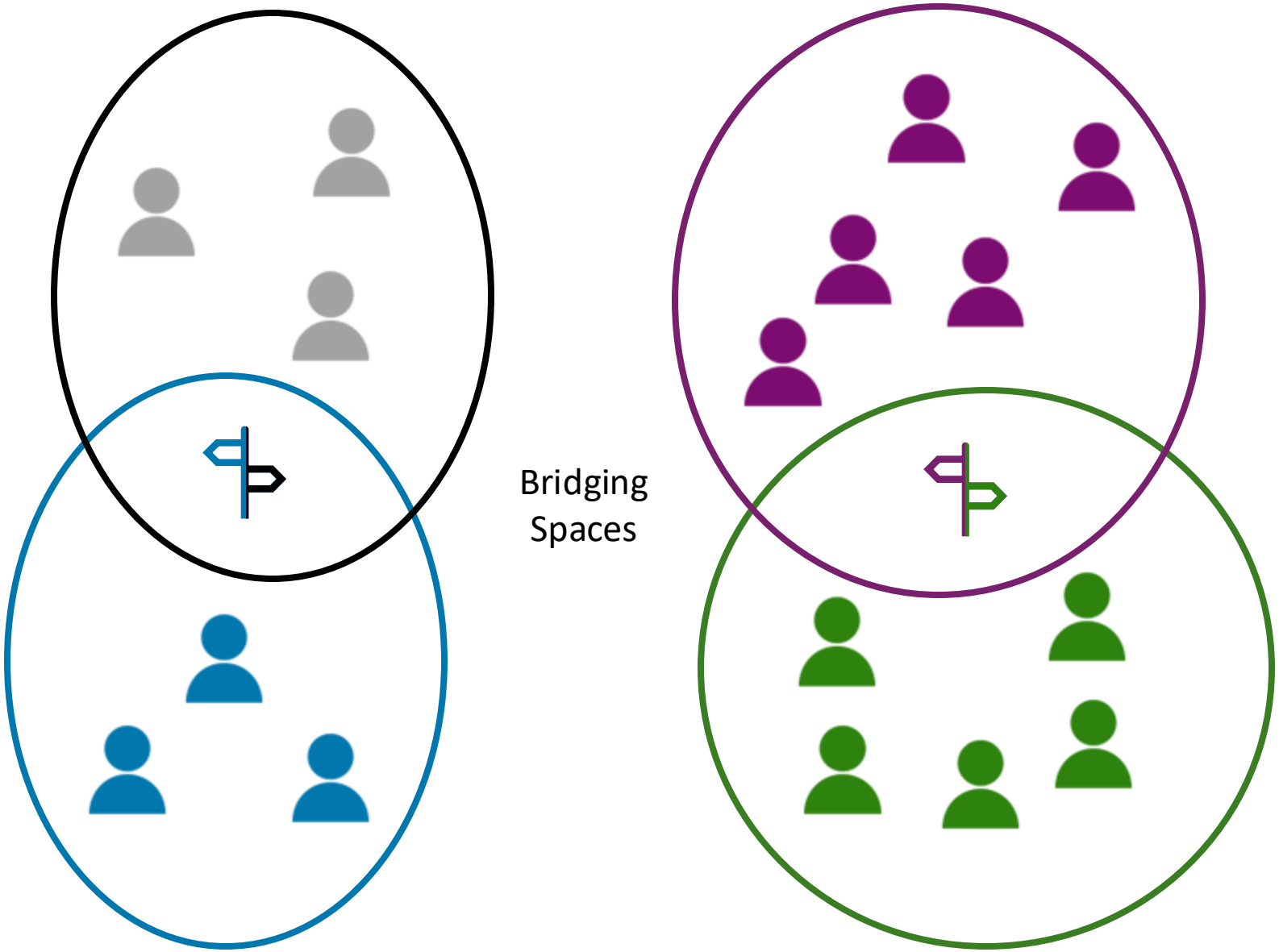
The problem of conflicting social spheres. Binder. (2009)

- ✓ Decentralization
- ✓ Community centered
- ✓ Empowering communities
- ✓ Self–governance

Decentralization ----------→ Isolating communities or increasing divides

Community centered ----------→ Discouraging cross-community connections





A pattern language. Alexander. (1977)

Exposure to opposing views on social media can increase polarization

Where are we?

"As of 2024, internet users spend more than six hours a day in online activities."

"Aggregation in homophilic clusters of users dominates online dynamics."

"Friends share substantially less news from opposing ideology."

Average daily time spent using the internet. DataReportal (2024)

The echo chamber effect on social media. Cinelli M et al. (2021)

Exposure to ideologically diverse news and opinion on Facebook. Eytan Bakshy et al. (2015)

Exposure to opposing views on social media can increase political polarization. Christopher A. Bail et al. (2018)

Algorithmic moderation and ranking define our experience online

We need to do something

"We need human moderators with cultural context, with meaningful accountability and transparency."

"Unilateral decisions are likely to be inconsistent, especially for marginalized communities."

"The work from these invisible hands forms the bedrock of how our internet was built and is regulated."

Content moderation in under-resourced regions. Tech Global Institute. (2023)

Black in moderation. Anika Collier Navaroli. (2023)

What we know about using non-engagement signals in content ranking. Tom Cunningham et al. (2024)

Moving towards more transparent community-centered approaches

Current opportunities

- ✓ Since 2019
- ✓ Decentralization
- ✓ Content moderation
- ✓ Algorithm customization

- ✓ Since 2008
- ✓ Decentralization
- ✓ Content moderation
- ✓ Community-specific rules

Where can researchers test algorithms?

- ✓ Algorithmic experimentation
- ✓ Decentralization
- ✓ Explainable moderation
- ✓ Surfacing differences
- ✓ Bridging communities

My dissertation

Research Questions

RQ 1. How can we design opportunities for community-centered, explainable decentralized governance?

RQ 2. How can we design bridging opportunities for communities with differing norms and values?

Contributions

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

Contributions

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

- ✓ Demonstrating how **community purpose and norms set the tone for pro-social discourse**.
- ✓ Providing large scale analysis of Reddit norms, mapping them to an **empirical schema of norms** that captures their nuances.
- ✓ Releasing **comprehensive dataset of over 230,000 norm-violating posts**, complete with moderation explanations.

Contributions

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

- ✓ Framing content moderation as an explainable classification task, releasing **the first rationale focused dataset for norm-violating content**, along with benchmarked NLP models.
- ✓ Defining **two new tasks to identify variations in norms interpretation** across communities, enabling mutual understanding and bridging governance gaps.

Contributions

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

- ✓ Designing and deploying, **Odessa, an open-source experimental social network** as a test space for human-AI interaction focused on decentralized governance.
- ✓ **Designing and evaluating bridging mechanisms** that enable cross-community interactions.

Contributions

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

Odessa
Social Network App

Contributions

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

What does each community care about?

Your community. Your values.

Research Questions:

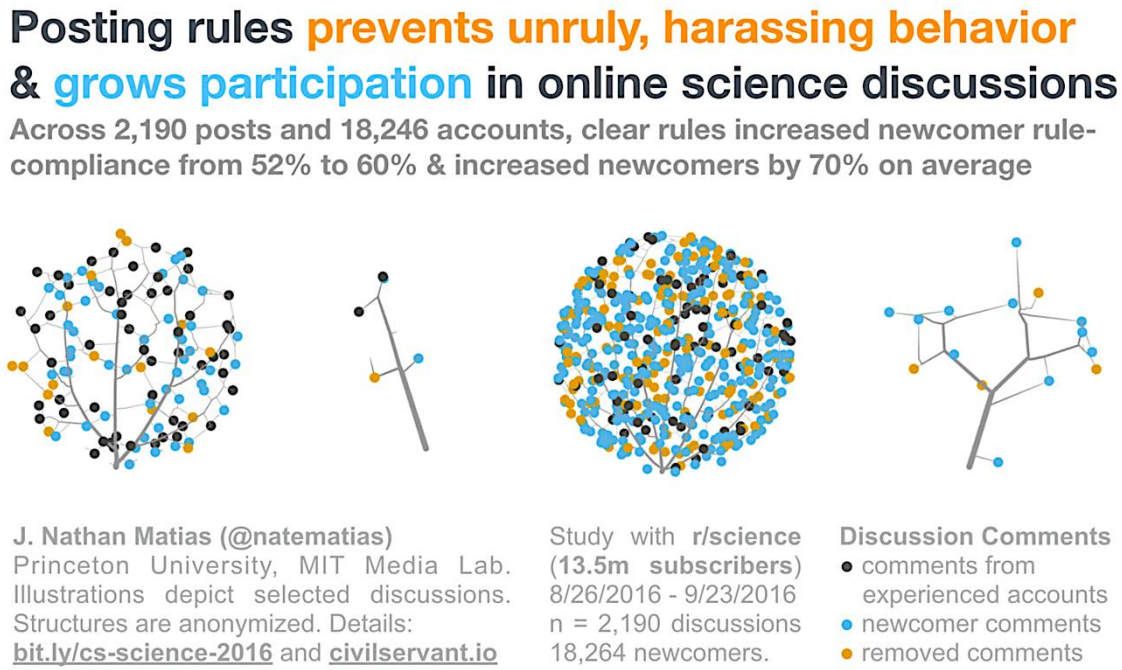
- What do communities care about?
- How do they express these values?

Evaluating hypothesis:

✓ Purpose

✓ Speech norms

Posting community rules prevents harassing behavior



Preventing harassment and increasing group participation. J. Nathan Matias (2019)

Releasing new dataset

Surfacing patterns from online communities

Reddit

+ 230K removed comments with moderators' follow-ups

+19K communities

+122K moderators

+ 3M removed comments

+ 60K unique rules

Real time tracking of original content

Mapped norms to removed content

Sinclair Target | Collaborator

Releasing new dataset

Surfacing patterns from online communities

Reddit

+ 230K removed comments with moderators'

follow-ups

Disclaimer
All examples were sampled from actual online content. Many of these examples include moderated content, a.k.a, **profanity**.

Community purpose sets the tone

COMPASSION	CURIOSITY	SELF STORY	PURPOSE	NAME
1.05	0.94	1.07	Buy, sell, trade! Upland is an open market, property trading game where the virtual properties are truly yours.	r/UplandMe
1.05	1.04	1.05	All about the most isolated city in the world and the fabulous people who live there.	r/perth
0.94	1.05	0.97	For everything to do with Forgeworld and the Horus Heresy. This includes 30k and 30k vs 40k battle reports; army list files, etc.	r/Warhammer30k
1.03	1.06	0.91	A place for people who enjoy carp fishing or want to learn about it!	r/CarpFishing

Perspective API. (2024)

Community purpose sets the tone

COMPASSION	CURIOSITY	SELF STORY	PURPOSE	NAME
1.58	1.61	1.70	This is a place for victims of narcissistic abuse to come together to support, encourage, learn from, share with, and validate.	r/NarcissisticAbuse
1.79	1.65	1.80	Hoarding disorder occurs in an estimated 2 to 6 percent of the population and often leads to substantial distress and problems.	r/ChildofHoarder
1.44	1.54	1.45	A place where you can ask veterinary medicine-related questions and get advice from veterinary professionals.	r/AskVet
1.50	1.46	1.47	This is a platform designed to inform and unite the NP community. Asking for advice, practice information, the job market, etc.	r/nursepractitioner
1.05	0.94	1.07	Buy, sell, trade! Upland is an open market, property trading game where the virtual properties are truly yours.	r/UplandMe
1.05	1.04	1.05	All about the most isolated city in the world and the fabulous people who live there.	r/perth
0.94	1.05	0.97	For everything to do with Forgeworld and the Horus Heresy. This includes 30k and 30k vs 40k battle reports; army list files, etc.	r/Warhammer30k
1.03	1.06	0.91	A place for people who enjoy carp fishing or want to learn about it!	r/CarpFishing
1.06	1.00	0.98	News and discussion about the Tor anonymity software. New to Tor? Please read the Tor FAQ!	r/TOR
0.68	0.68	0.70	Welcome to r/TheBidenShitShow This is a sub to watch and discuss the mumblings and bad decisions of the Joe Biden presidency.	r/TheBidenShitshow
0.74	0.69	0.75	Paintings and drawings of barren, overgrown, devastated or post-apocalyptic landscapes and the characters, tech, and monsters within.	r/ImaginaryWastelands
0.73	0.62	0.65	When a good meme is ruined by a shitty caption.	r/comedyhomicide
0.66	0.63	0.68	Become a ConservativeMemes subscriber! Post your Conservative Memes later at r/ConservativeMemes.	r/ConservativeMemes

Perspective API. (2024)

Community purpose sets the tone

COMPASSION	CURIOSITY	SELF STORY	PURPOSE	NAME
1.58	1.61	1.70	This is a place for victims of narcissistic abuse to come together to support, encourage, learn from, share with, and validate.	r/NarcissisticAbuse
1.79	1.65	1.80	Hoarding disorder occurs in an estimated 2 to 6 percent of the population and often leads to substantial distress and problems.	r/ChildofHoarder
0.74	0.69	0.75	Paintings and drawings of barren, overgrown, devastated or post-apocalyptic landscapes and the characters, tech, and monsters within.	r/ImaginaryWastelands
0.73	0.62	0.65	When a good meme is ruined by a shitty caption.	r/comedyhomicide
0.66	0.63	0.68	Become a ConservativeMemes subscriber! Post your Conservative Memes later at r/ConservativeMemes.	r/ConservativeMemes

Perspective API. (2024)

Community norms set the tone

Short Name	Description
sarcasm, cynicism, and profanity	Do not make posts that consist only of sarcasm, cynicism, or virtue signaling that do not add to the discussion. Profanity is not prohibited but is discouraged to facilitate discussion.
post scary stories	Please only post scary stories. This is r/scarystories — do not post questions, discussions, videos, images, etc.

Community norms set the tone

Short Name	Description
sarcasm, cynicism, and profanity	Do not make posts that consist only of sarcasm, cynicism, or virtue signaling that do not add to the discussion. Profanity is not prohibited but is discouraged to facilitate discussion.
post scary stories	Please only post scary stories. This is r/scarystories — do not post questions, discussions, videos, images, etc.
don't be a dick	If you're not going to behave yourself, you can go someplace else.
be nice	Be nice to other Reddit users.

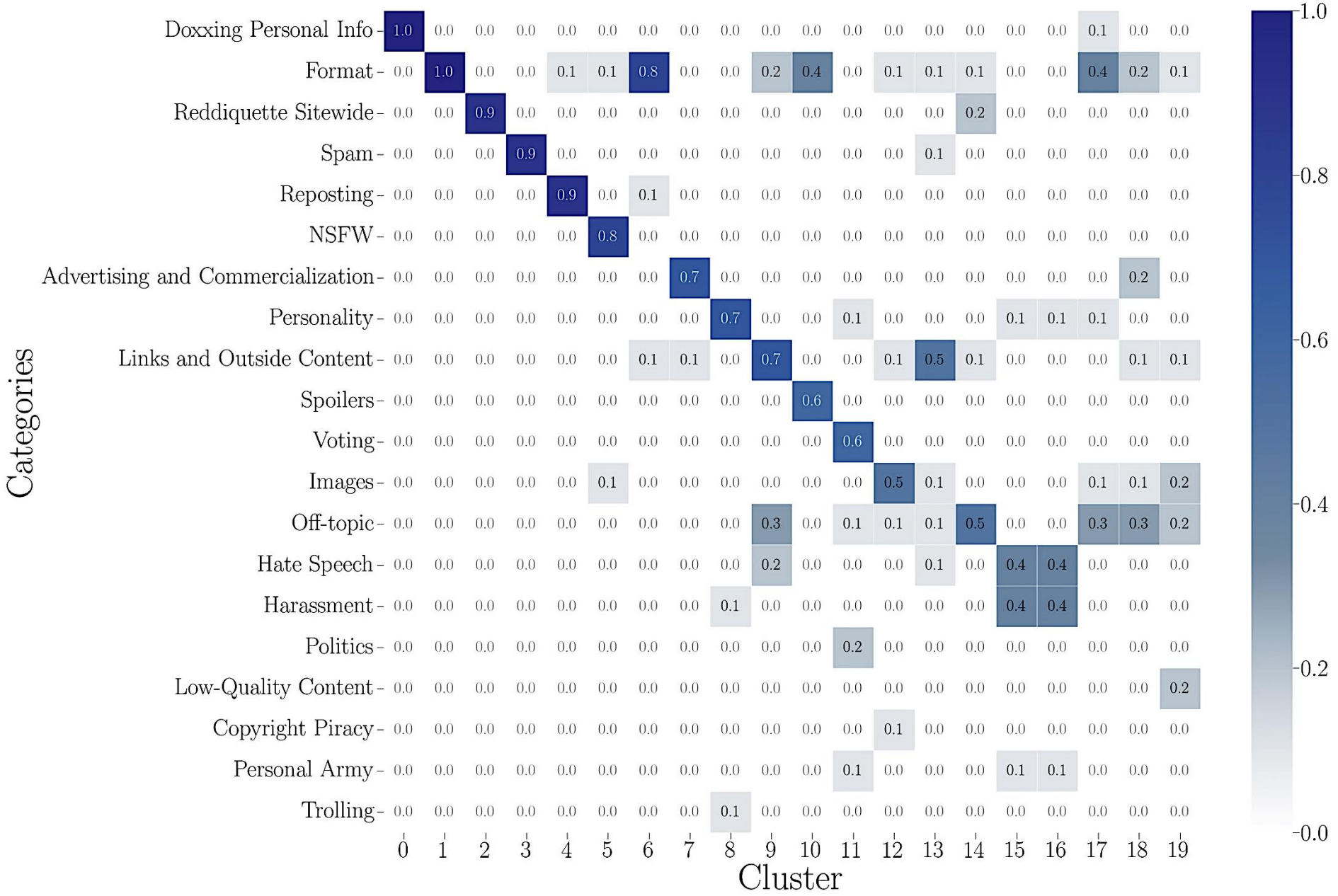
Community norms set the tone

Short Name	Description
sarcasm, cynicism, and profanity	Do not make posts that consist only of sarcasm, cynicism, or virtue signaling that do not add to the discussion. Profanity is not prohibited but is discouraged to facilitate discussion.
post scary stories	Please only post scary stories. This is r/scarystories — do not post questions, discussions, videos, images, etc.
don't be a dick	If you're not going to behave yourself, you can go someplace else.
be nice	Be nice to other Reddit users.
no sob/emotional string	We do not permit titles that are of a sob nature or pull.
all comments must be positive	All comments must be positive.
no married/partnered/poly	This sub is for single and foreveralone people only: [...]
no story posts	This is not a subreddit for stories. [...]
no medical advice	No posts asking for medical, injury, or pain-related advice.
be respectful and kind to one another	There is nothing wrong with disagreement or arguing, but keep it respectful and civil. Any instances of name-calling, hate speech, or unkind behavior will not be tolerated.

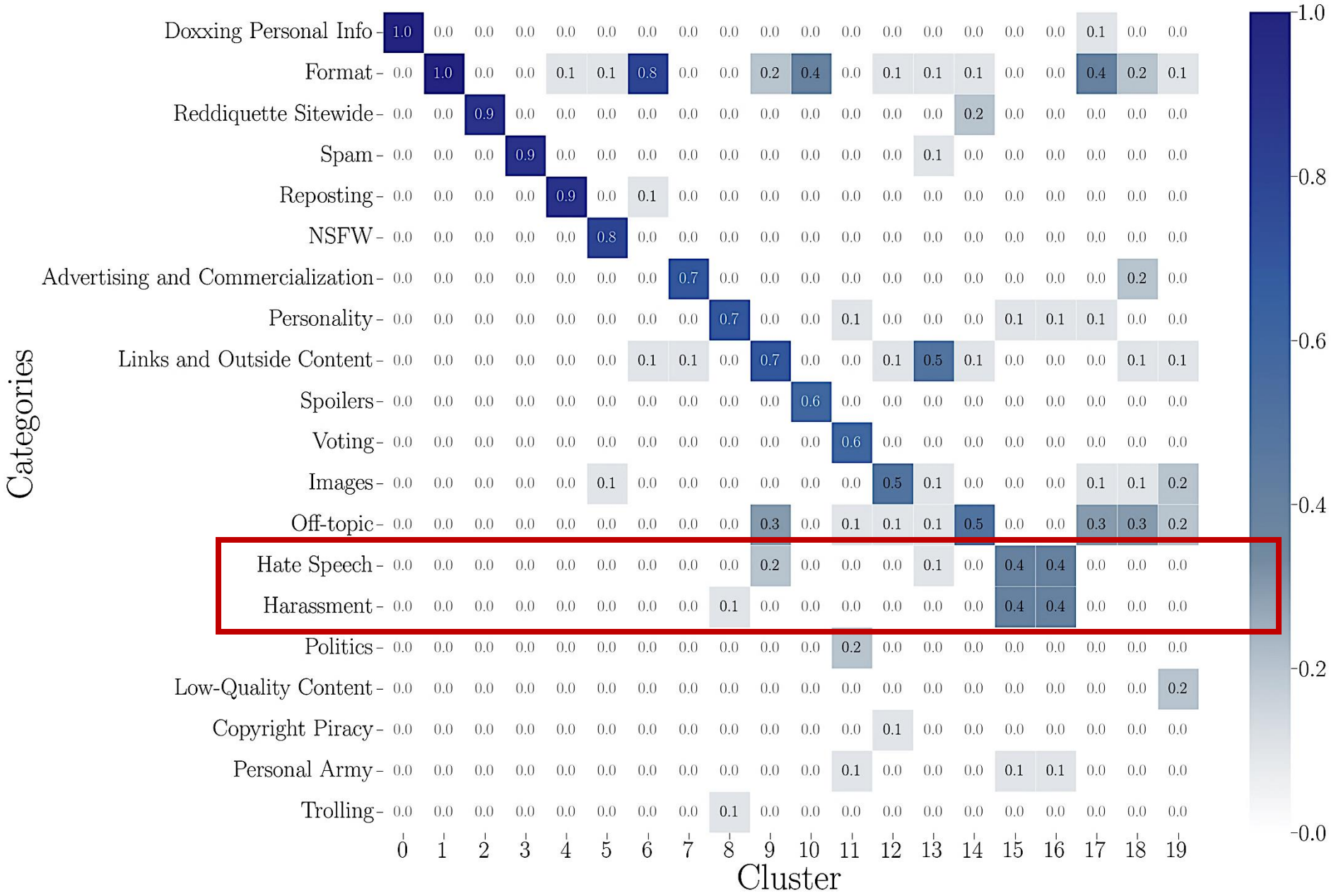
Community norms set the tone

Short Name	Description	Type
sarcasm, cynicism, and profanity	Do not make posts that consist only of sarcasm, cynicism, or virtue signaling that do not add to the discussion. Profanity is not prohibited but is discouraged to facilitate discussion.	**Restrictive**
post scary stories	Please only post scary stories. This is r/scarystories — do not post questions, discussions, videos, images, etc.	**Prescriptive**
don't be a dick	If you're not going to behave yourself, you can go someplace else.	**Restrictive**
be nice	Be nice to other Reddit users.	**Prescriptive**
no sob/emotional string	We do not permit titles that are of a sob nature or pull.	**Restrictive**
all comments must be positive	All comments must be positive.	**Prescriptive**
no married/partnered/poly	This sub is for single and foreveralone people only: [...]	**Prescriptive**
no story posts	This is not a subreddit for stories. [...]	**Restrictive**
no medical advice	No posts asking for medical, injury, or pain-related advice.	**Restrictive**
be respectful and kind to one another	There is nothing wrong with disagreement or arguing, but keep it respectful and civil. Any instances of name-calling, hate speech, or unkind behavior will not be tolerated.	**Prescriptive**

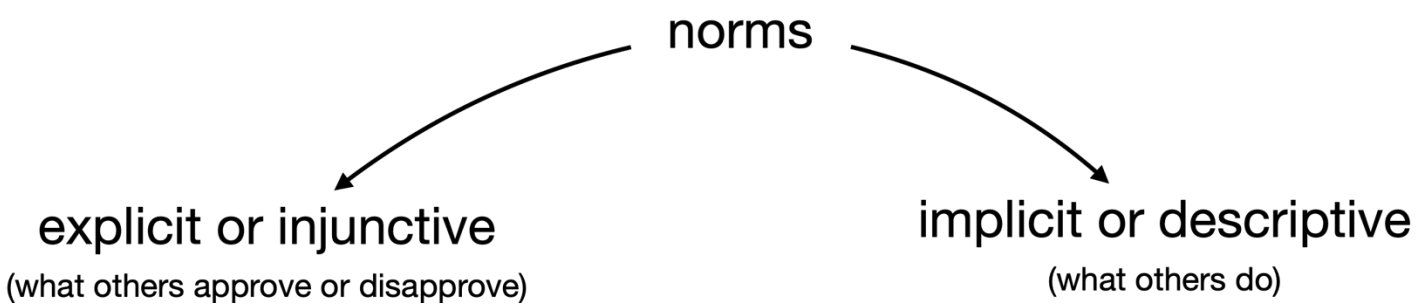
Some norms are better understood than others



Norms are not clearly differentiated across communities

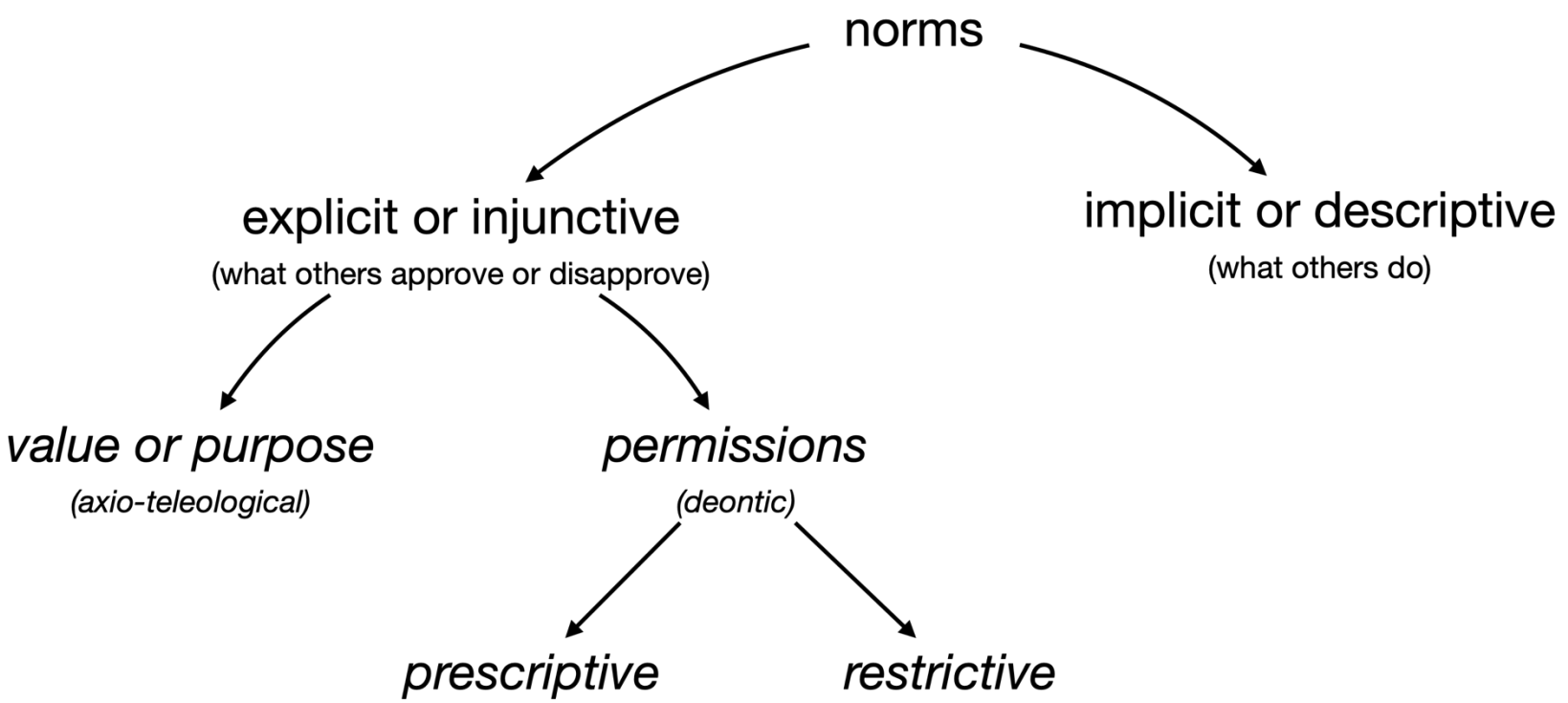


Speech norms taxonomy



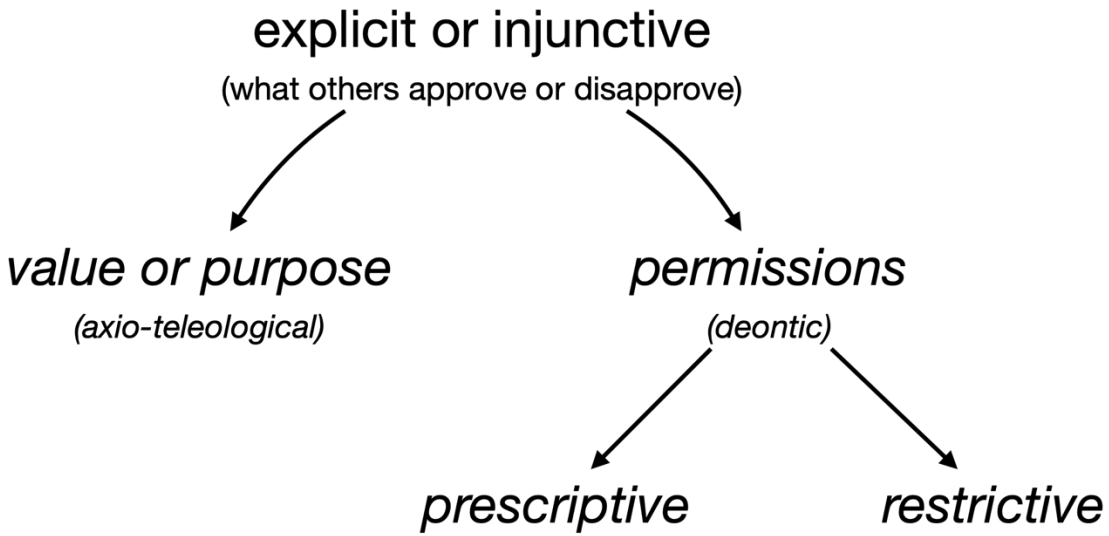
Normative Conduct. Cialdini. (1990)
Descriptive and Injunctive Norms in a Wikipedia. Morgan and Filippova. (2018)

Speech norms taxonomy



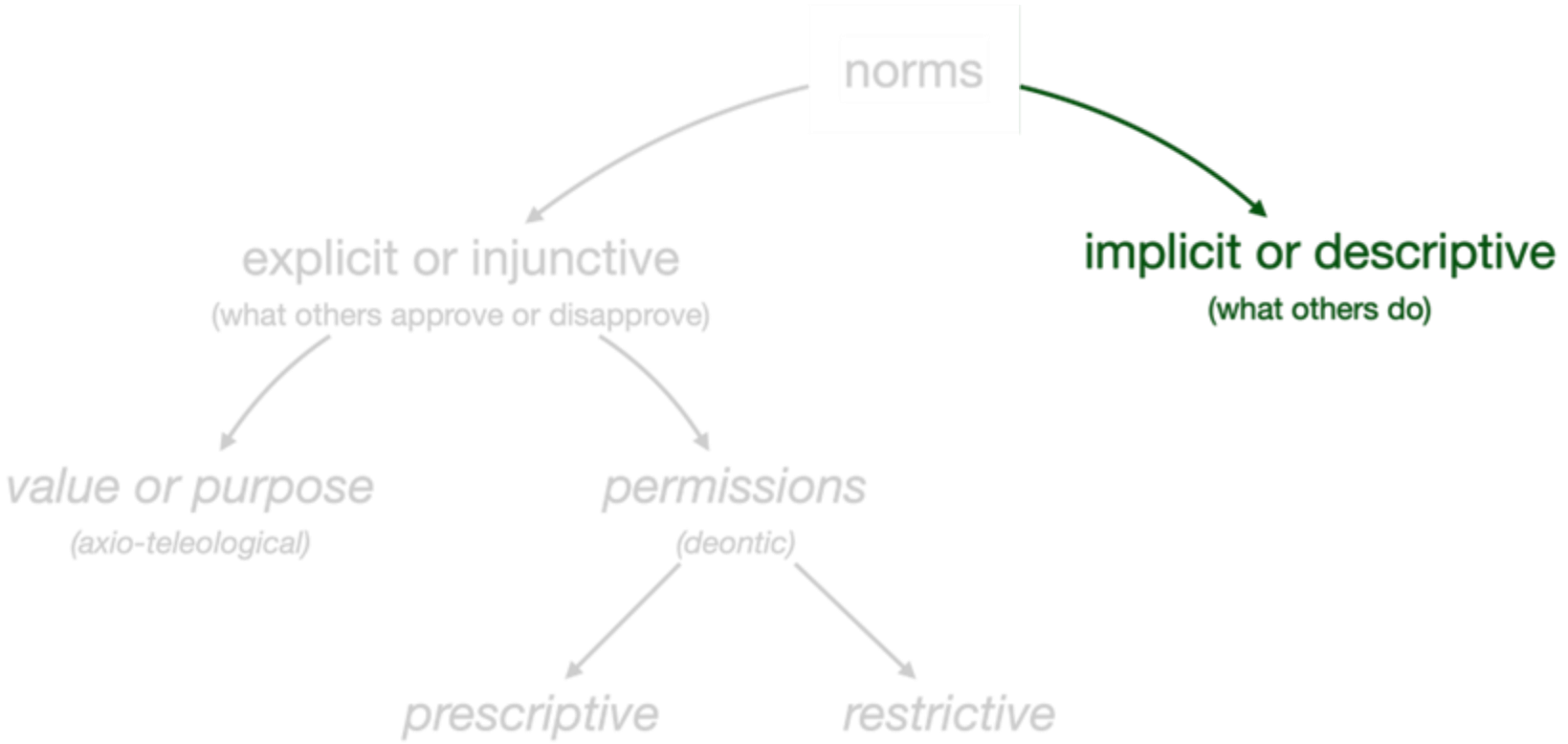
Normative Conduct. Cialdini. (1990)
Descriptive and Injunctive Norms in a Wikipedia. Morgan and Filippova. (2018)

Speech norms taxonomy



Normative Conduct. Cialdini. (1990)
Descriptive and Injunctive Norms in a Wikipedia. Morgan and Filippova. (2018)

Speech norms taxonomy



Normative Conduct. Cialdini. (1990)
Descriptive and Injunctive Norms in a Wikipedia. Morgan and Filippova. (2018)

Contributions

Surfacing patterns from online communities

- ✓ Release of unique dataset, largest to our knowledge, to study norms (+60k) and content moderation (+230K).
- ✓ Large scale analysis of pro-social metrics within online communities conditioned on community purpose.
- ✓ Design considerations for turning norms taxonomy into features in Odessa.

Contributions

Research Questions

RQ 1. How can we design opportunities for community-centered, explainable decentralized governance?

- ✓ Release of unique dataset, largest to our knowledge, to study norms (+60k) and content moderation (+230K).
- ✓ Large scale analysis of pro-social metrics within online communities conditioned on community purpose.
- ✓ Design considerations for turning norms taxonomy into features in Odessa.

What does each community care about?

Your community. Your values.

Research Questions:

- What do communities care about?
- How do they express these values?

Evaluating hypothesis:

✓ Purpose

✓ Speech norms

What does each community care about?

Your community. Your values.

Research Questions:

- What do communities care about?
- How do they express these values?
- **How can we operationalize these values? ➜ Tools**

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

Part I: Data

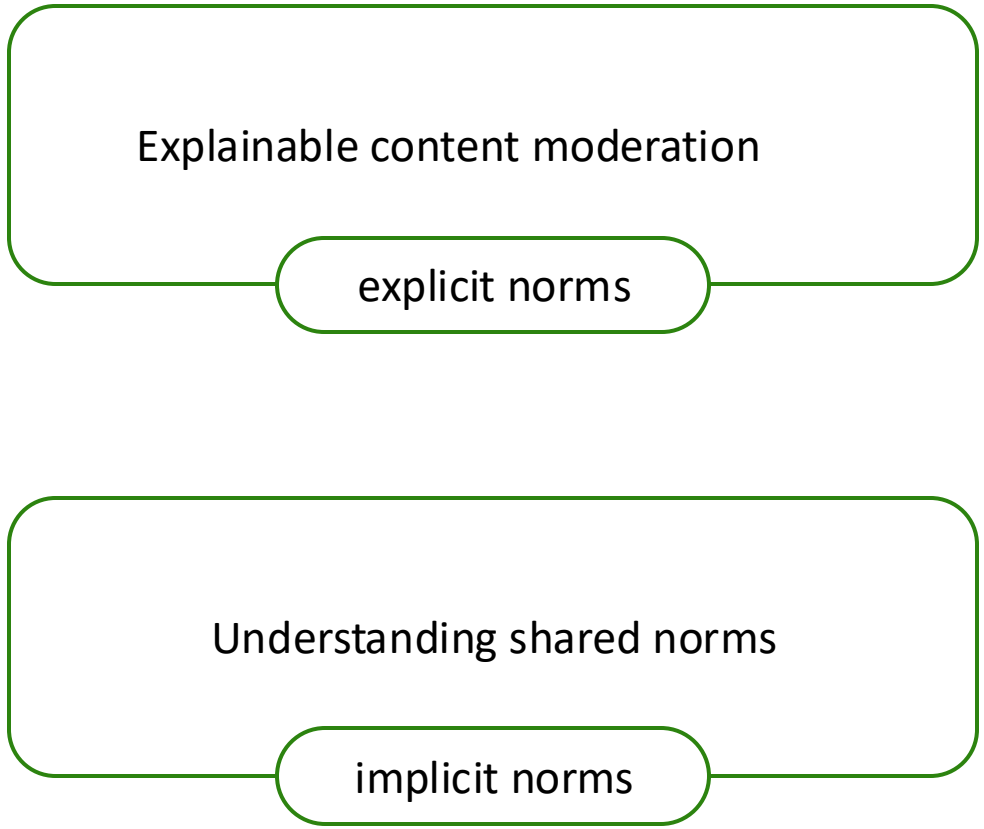
Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App



Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

Explainable content moderation

explicit norms

Sasha Rush | Collaborator

Understanding shared norms

implicit norms

Assisting explainable, decentralized governance

"We need human moderators with cultural context."

"Unilateral content policies are likely to be inconsistent, especially for marginalized communities."

Content moderation in under-resourced regions. Tech Global Institute. (2023).

"Sadly, we do get a lot of burnout, […]. As far as keeping everything alive, we just keep moving forward. Try and encourage them. But, you know, it's tough."

Moderation practices as emotional labor in sustaining online communities. Dosono, B., & Semaan, B. (2019).

Moderation is taxing and difficult

“You literally ran from your house and exhibit signs of homesickness for it. Lmao the world would not miss your absence, in fact your existence is nuisance. Do you know where you are? You are not welcome here. Like at all.”

Discouraged behavior: **don’t be a dick.**

Moderation benefits from A.I. assistance

"You literally ran from your house and exhibit signs of homesickness for it. Lmao the world would not miss your absence, in fact your existence is nuisance. Do you know where you are? You are not welcome here. Like at all."

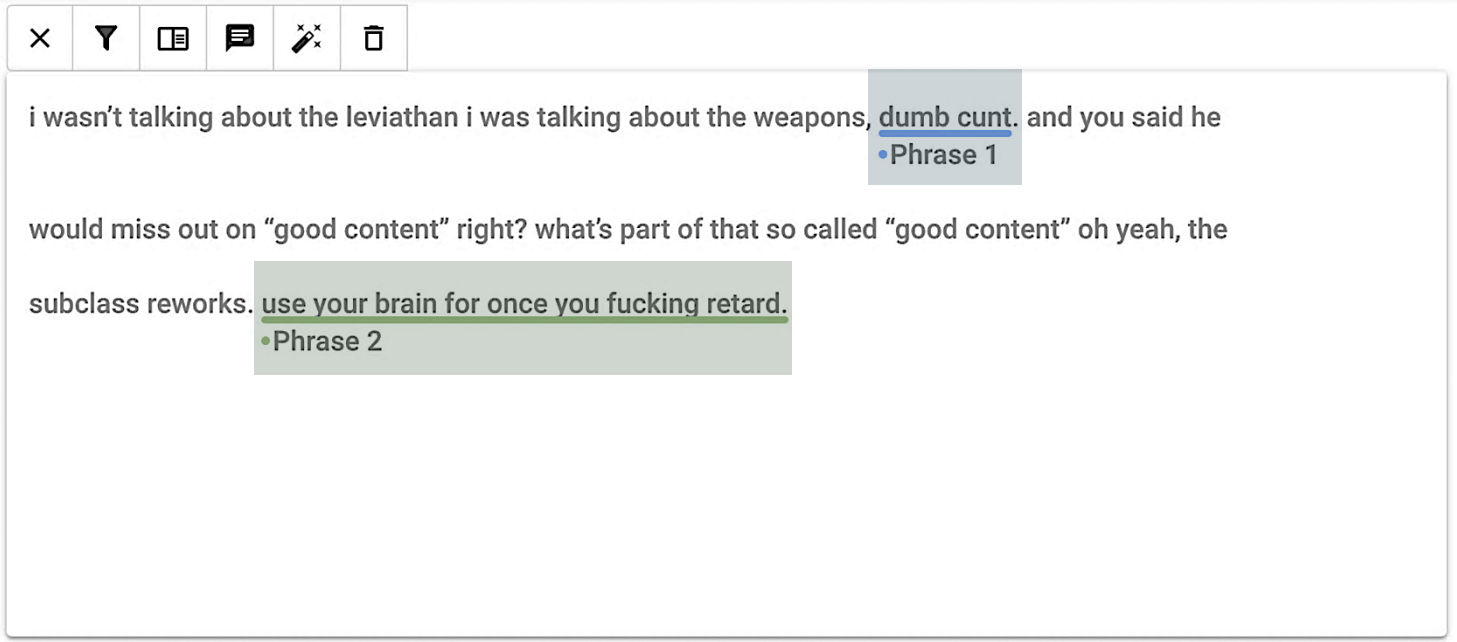
Discouraged behavior: **don't be a dick.**

Framing moderation as explainable classification task

NLP Task

Given a comment, extract the exact the violating subsequence that triggers the violation

Collecting human-generated rationales



Crowdsourcing task

- ❑ 200 comments
- ❑ 4 annotators each*
- ❑ lengths between 200 and 400 characters
- ✓ **Result**: extracted around 20% of text as explanation

Explainable and scalable approaches

- Integrated gradients
- Saliency
- Lime
- Discrete rational

BERT. Devlin at al., (2018) Deep inside convolutional networks. Simonyan et al. (2013) Why should I trust you? Ribeiro et al. (2016)

Axiomatic attribution for deep networks. Sundararajan et al. (2017) Rationalizing neural predictions. Lei et al. (2016)

Explainable and scalable approaches

- Integrated gradients
- Saliency
- Lime
- **Discrete rational**

BERT. Devlin at al., (2018)

Deep inside convolutional networks. Simonyan et al. (2013)

Why should I trust you? Ribeiro et al. (2016)

Axiomatic attribution for deep networks. Sundararajan et al. (2017)

Rationalizing neural predictions. Lei et al. (2016)

Prompting moderators and language models for explainable decisions

Instructions, select the violating portion of this comment:

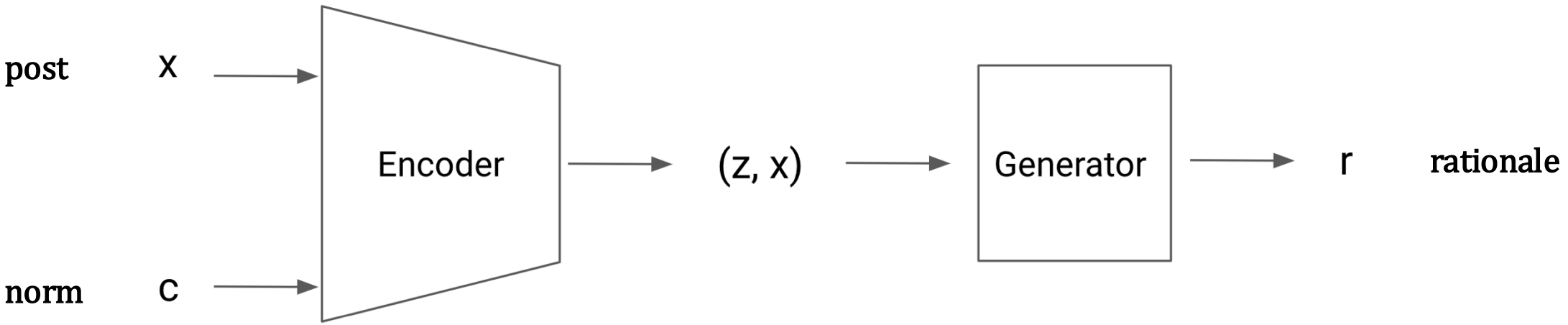
- It is a collection of up to three phrases, where each phrase is as short as possible.
- Each phrase provides sufficient evidence for triggering norm violation.
- If the collection of phrases was removed, the comment wouldn't violate the norm.

Prompting moderators and language models for explainable decisions

Instructions, select the violating portion of this comment:

- It is a collection of up to three phrases, where **each phrase** is **as short as possible**.
- **Each phrase provides sufficient evidence** for triggering norm violation.
- If the collection of phrases was removed, the comment wouldn't violate the norm.

High-level approach



$$z \sim enc(x, c) \equiv p(z|x, c)$$

Rationalizing neural predictions. Lei, T., Barzilay, R., & Jaakkola, T. (2016)

High-level approach

$$\text{enc}(\text{ gen}(\text{x})\ ) = \text{enc}(\text{ z, x }) \approx \text{enc}(\text{x}) = \tilde{\text{y}}$$

Joint objective

1. The rationale must **suffice**:

$$\mathcal{L}(\mathbf{z}, \mathbf{x}, \mathbf{y}) = \left\| \mathbf{enc}(\mathbf{z}, \mathbf{x}) - \mathbf{y} \right\|_2^2$$

2. **Short** and **coherent** rationales:

$$\Omega(\mathbf{z}) = \lambda_1 \|\mathbf{z}\| + \lambda_2 \sum_t |\mathbf{z}_t - \mathbf{z}_{t-1}|$$

$$\min_{\theta_e, \theta_g} \sum_{(\mathbf{x},\mathbf{y}) \in D} \mathbb{E}_{\mathbf{z} \sim \mathbf{gen}(\mathbf{x})} \left[\text{cost}(\mathbf{z}, \mathbf{x}, \mathbf{y}) \right]$$

Rationalizing neural predictions. Lei, T., Barzilay, R., & Jaakkola, T. (2016)

Explainable and scalable approaches

TABLE I
$P(y|w_{1:l})$

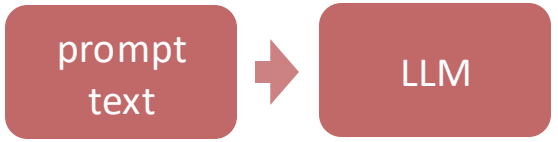
Model	F1-score
BERT+LSTM - Attention	.86
BERT+LSTM - Gradient	.86
BERT+LSTM - Lime	.86
Bert-to-Bert (Lei et al.'s)	.83

TABLE II
BLEU SCORE. INTER-ANNOTATOR AGREEMENT.

BLEU	.63
P-1	.73
P-2	.66
P-3	.60
P-4	.54

(Saldías et al. 2020)

Large language models can help!



BERT-based approaches

- Require post processing
- Require some fine tuning
- As norms change and evolve these may become less generalizable

LLM-based approaches

- Prompt takes care of processing
- No need for fine tuning
- More easily generalizes with few examples

CoPE, the Content Policy Evaluator
Hoover and the Cyber Policy Center discuss tech policy at Stanford.
Michael McFaul and Amy Zegart. (2024).

Language models are few-shot learners. Brown at al. (2020)

Contributions

Self-governance and surfacing differences in shared norms

- ✓ Re-imagined content moderation as an explainable task.
- ✓ Benchmarked interpretable BERT-based and compared to human performance.
- ✓ Collected human annotated explanations for hundreds of moderated posts for future evaluations.

Contributions

Research Questions

RQ 1. How can we design opportunities for community-centered, explainable decentralized governance?

- ✓ Re-imagined content moderation as an explainable task.
- ✓ Benchmarked interpretable BERT-based and compared to human performance.
- ✓ Collected human annotated explanations for hundreds of moderated posts for future evaluations.

Part I: Data

Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

Explainable content moderation

explicit norms

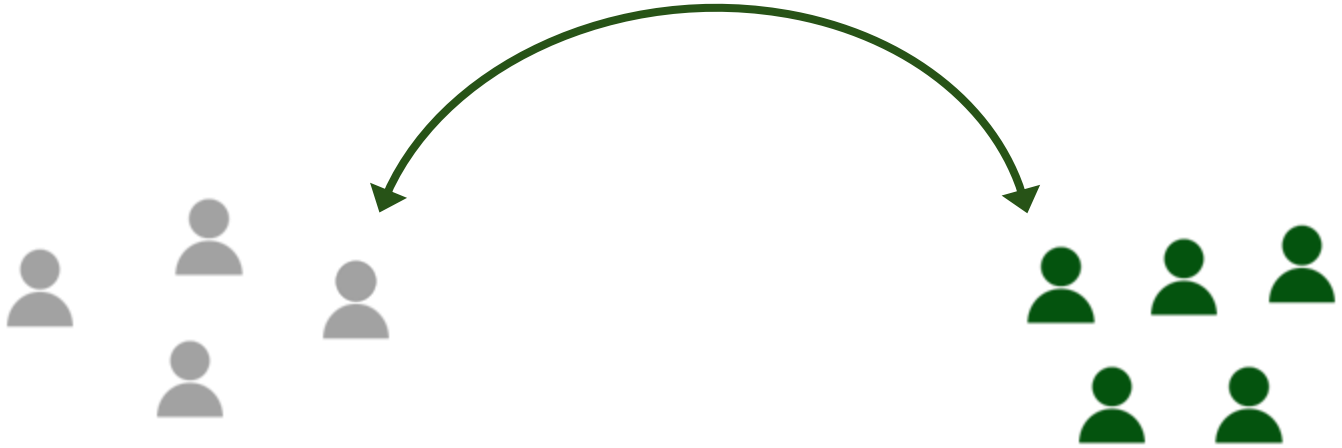
Understanding shared norms

implicit norms

Sasha Krigel |
Collaborator, Student

Understanding shared norms

Creating bridges



Understanding shared norms

Task 1

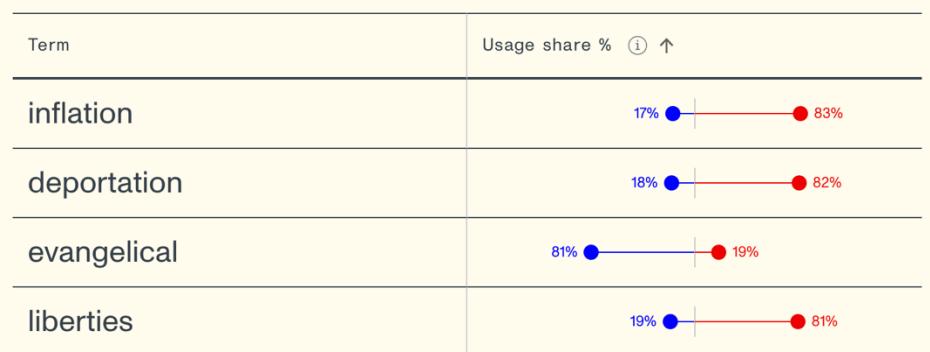
How can we **surface shared norms** among communities?

Task 2

How can we uncover **underlying differences for the same norm**?

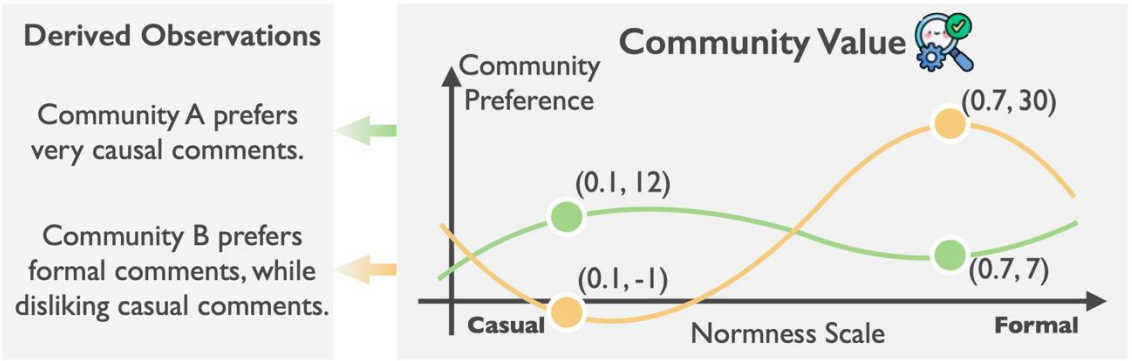
Looking at differences across communities

Word-usage level – Political parties



Bridging Dictionary: AI-Generated Dictionary of Partisan Language Use. Hang Jiang et al. (2024)

Community level – Six categories of norms



ValueScope: Unveiling Implicit Norms and Values via Return Potential Model of Social Interactions. Chan Y. Park et al. 2024.

Finding differences for the same norms

Our approach

Norm: Disrespectful Personal Attacks

Community 1 (r/medicine):

Comments that aim to demean or **belittle individuals based on their views** are not tolerated.

Community 2 (r/Texas):

Disrespect is identified as **belittling remarks** about **a person's intelligence or character**.

Two unsupervised tasks

Our approach

Task 1

- ❑ Identify shared norms and their community-specific interpretations

Task 2

- ❑ Uncover how *the same* norm is enforced differently across two communities

Two unsupervised tasks

Our approach

Task 1

- ❑ Identify **shared norms and their community-specific interpretations**

Task 2

- ❑ Uncover how *the same* norm is enforced differently across two communities

Pipeline to compare communities

Our approach

Sample N comments from each of two community → Prompt LLM → Response must cite comments from each community → Quantitative Metrics; Qualitative Analysis

Pipeline to compare communities

Our approach

Sample N comments from each of two community

Prompt LLM

Response must cite comments from each community

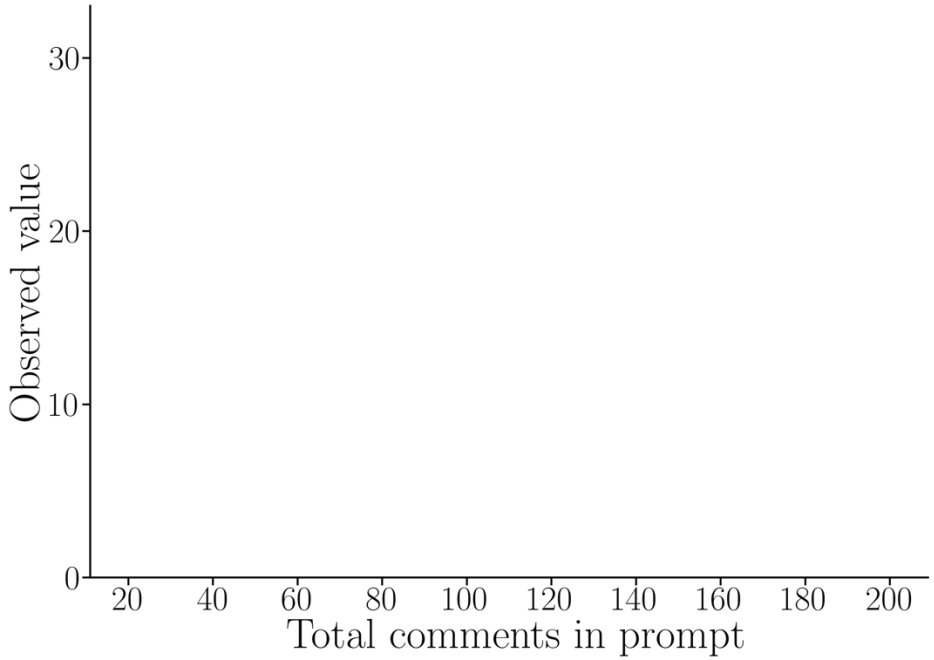

```
{
    "comm_a": "t5_2qnkr",
    "comm_b": "t5_2r05i",
    "name": "Please Stay Respectful",
    "description": "Respectful dialogue encourages constructive [...]",
    "comm_a_desc": "Community A views respect as crucial for productive discussions [...]",
    "comm_b_desc": "Community B emphasizes a more relaxed understanding of respect [...]",
    "comm_a_comment_ids": [
        2169,
        12589,
        20429
    ],
    "comm_b_comment_ids": [
        12325,
        13544,
        13962
    ]
}
```


Quantitative Metrics

Qualitative Analysis

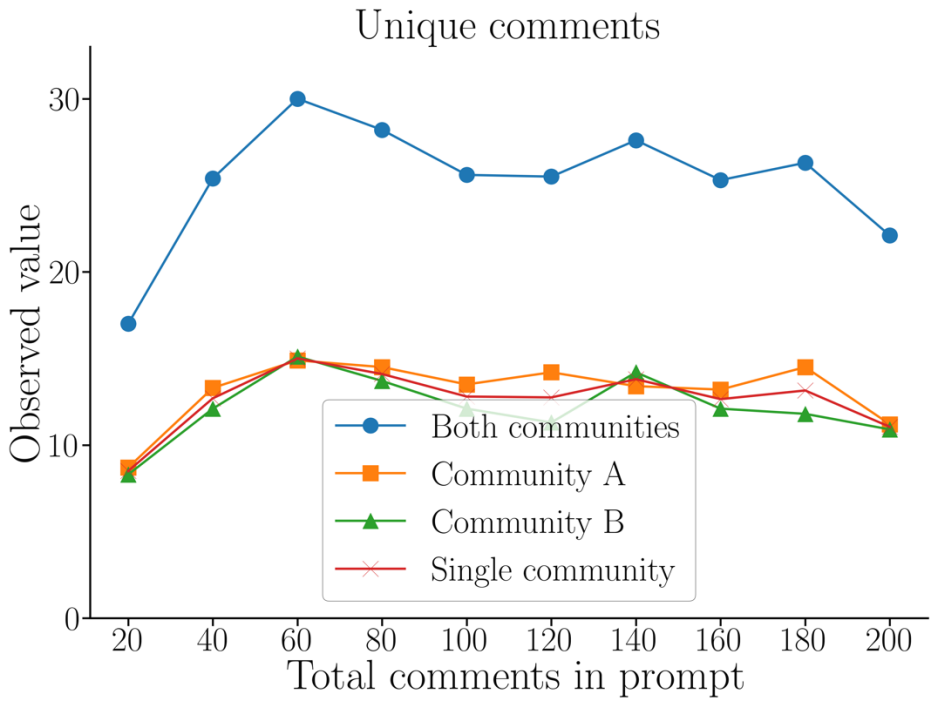
Requesting implicit norms

Task 1



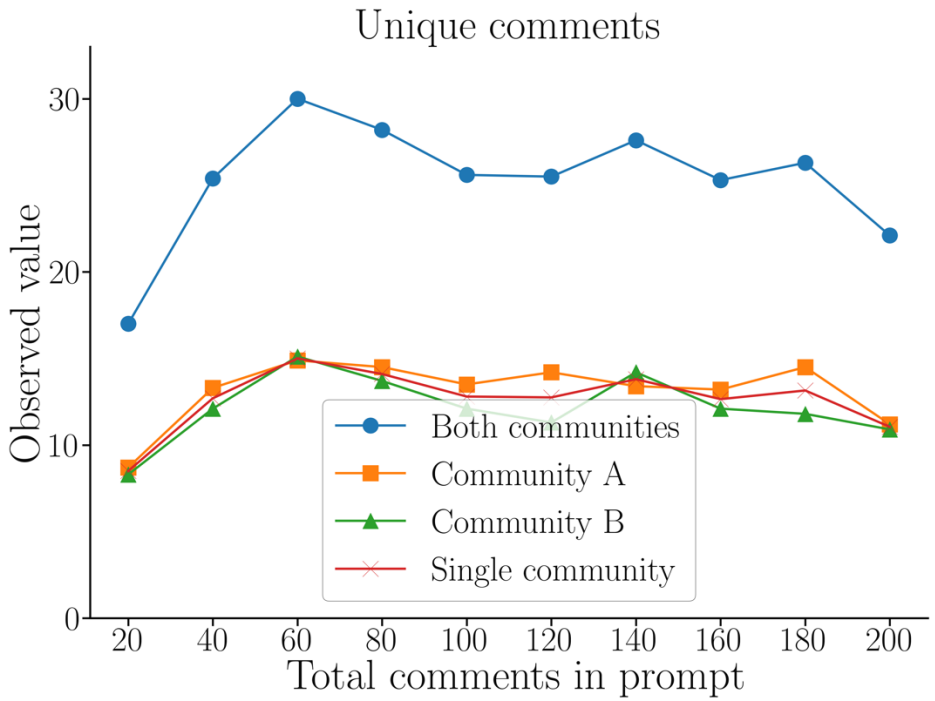
Communities are represented equally

Task 1



Communities are represented equally

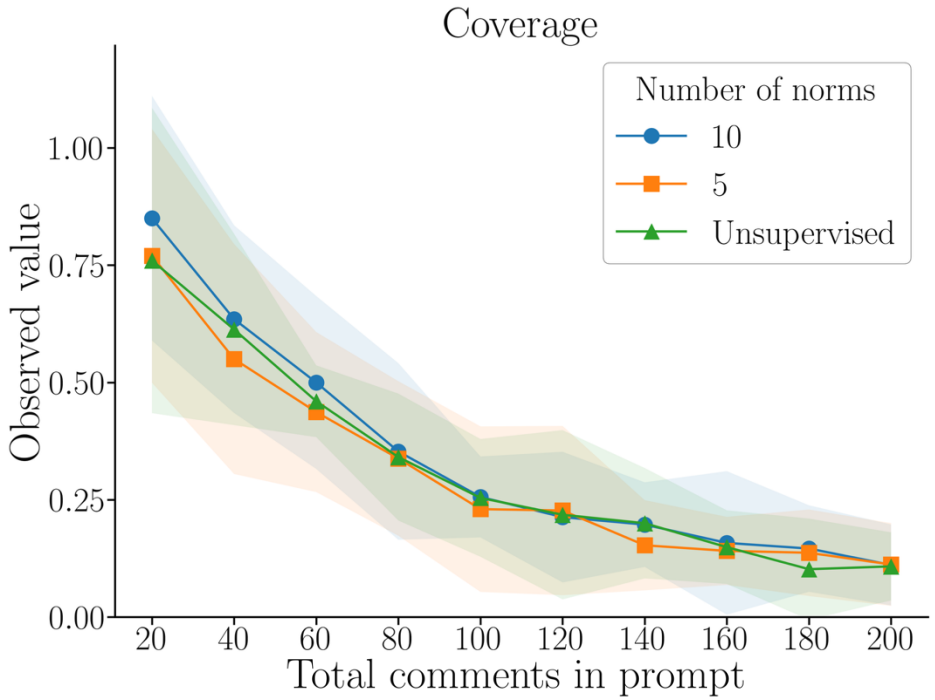
Task 1



- ✓ Run a systematic series of prompt-engineering improvements
- ✓ LLM's output is as expected and stable
- ✓ Evaluated convergence and performance of:
  - ✓ Coverage
  - ✓ Redundancy
  - ✓ Community representation
  - ✓ Number of norms generated
  - ✓ Proportion of violating comments
  - ✓ Adjusted for comment length

Communities are represented equally

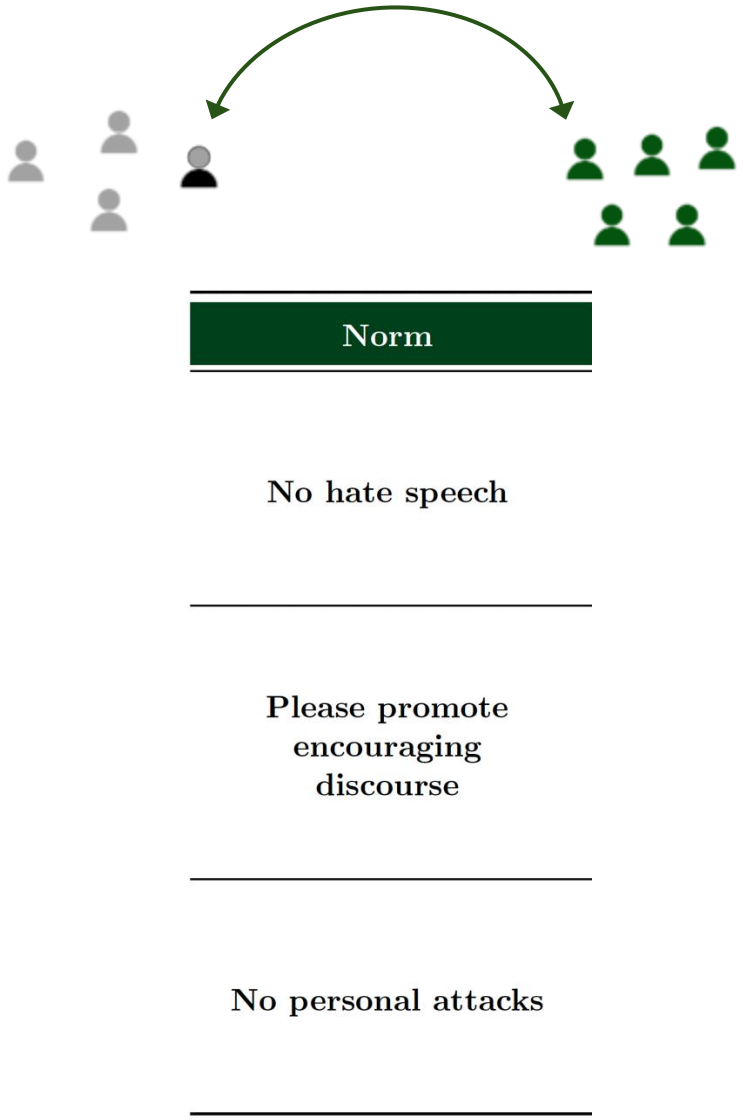
Task 1



- ✓ Run a systematic series of prompt-engineering improvements
- ✓ LLM's output is as expected and stable
- ✓ Evaluated convergence and performance of:
  - ✓ Coverage
  - ✓ Redundancy
  - ✓ Community representation
  - ✓ Number of norms generated
  - ✓ Proportion of violating comments
  - ✓ Adjusted for comment length

Shared implicit norms

Making implicit norms explicit



- ✓ Effectively surface implicit norms with different interpretations among two communities.
- ✓ We can use this method to increase awareness differences among communities.
- ❑ How are these different per community?

→ Task 2

Two unsupervised tasks

Task 2

Task 1

- ❑ Identify shared norms and their community-specific interpretations

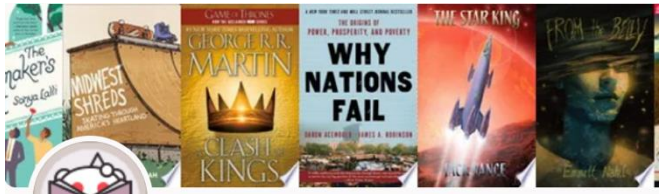
Task 2

- ❑ Uncover how ***the same*** **norm is enforced differently** across two communities

Generating definition per community

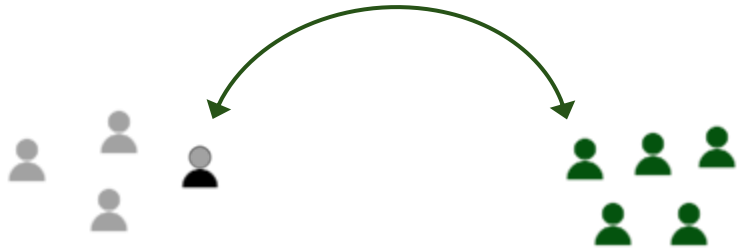
Norm	Definition (Community A)	Definition (Community B)
Be respectful	Community A emphasizes respectful dialogue, constructive criticism, and the importance of presenting opinions in a calm and logical manner. Disrespect is defined by the use of offensive language, personal attacks, or dismissive tones that undermine the validity of others' perspectives. Comments that fail to engage meaningfully with the topic and resort to name-calling or condescension are considered disrespectful.	Community B focuses on professional discourse, particularly in the context of medical practice and healthcare. Disrespect is characterized by a lack of professionalism, belittling attitudes towards healthcare providers or patients, and failure to recognize the emotional and knowledgeable contributions of others. This community values nuanced discussions and specific knowledge, so broad generalizations and inflammatory statements are deemed disrespectful.

Revealing accurate interpretations

Norm	Definition (Community A)	Definition (Community B)
Be respectful	 r/books Rule 2: Please use a civil tone and assume good faith when entering a conversation. Please conduct yourself as though in a family-friendly environment. Do not use obscenities, slurs, or gender-charged insults.	 r/medicine r/medicine is a public forum that represents the medical community and comments should reflect this. Please keep disagreement civil and focused on issues. Trolling, abuse, and insults (either personal or aimed at a specific group) are not allowed. Do not attack other users' flair.

- ✓ Unsupervised approach
- ✓ Small sample of violating comments needed
- ✓ Approach generalizes to other tasks

Why prompting two communities?



Joint Prompt

Independent Prompts

Definitions are statistically more specific when the model has access to outside perspectives

	Cosine Similarity	
	Definitions	**Cited Comments**
Joint Prompt	0.85 ± 0.08	0.51 ± 0.11
Independent Prompts	0.90 ± 0.09	0.51 ± 0.12
p-value (two-sample t-test)	$6.6e-10^{***}$	0.032

- ✓ Cited comments are not statistically different to each other

Future opportunities

Self-governance and **surfacing differences in shared norms**

- ❑ Surfacing commonalities and differences among communities of interest.
- ❑ Assisting content policy development.

Contributions

Self-governance and surfacing differences in shared norms

- ✓ Definition of two new NLP tasks with quantitative and qualitative benchmarks.
- ✓ Method to identify norm variations and foster mutual understanding.
- ✓ Systematic evaluation framework to evaluate models comparing community norms.

Contributions

Research Questions

RQ 2. How can we design bridging opportunities for communities with differing norms and values?

- ✓ Definition of two new NLP tasks with quantitative and qualitative benchmarks.
- ✓ Method to identify norm variations and foster mutual understanding.
- ✓ Systematic evaluation framework to evaluate models comparing community norms.

Part I: Data

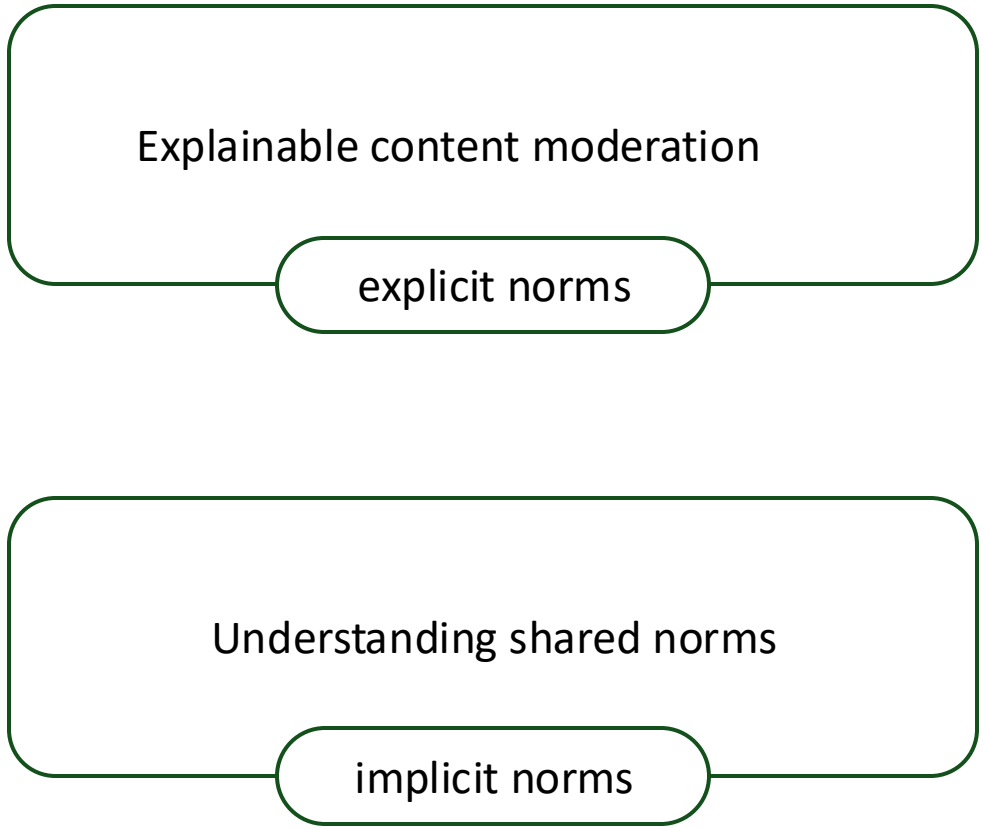
Surfacing patterns from online communities

Part II: Tools

Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App



Part I: Data

Surfacing patterns from online communities

Part II: Tools

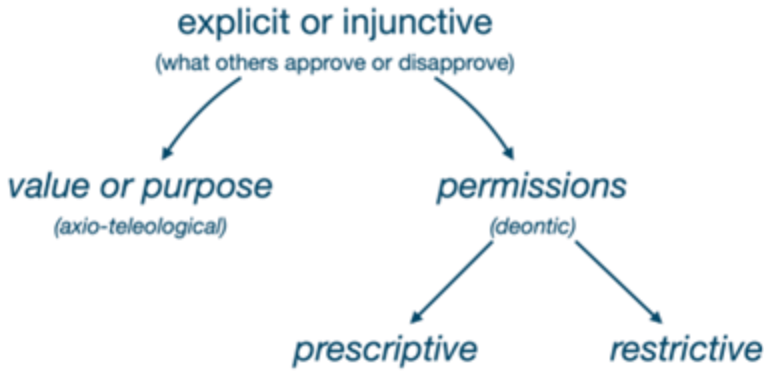
Self-governance and surfacing differences

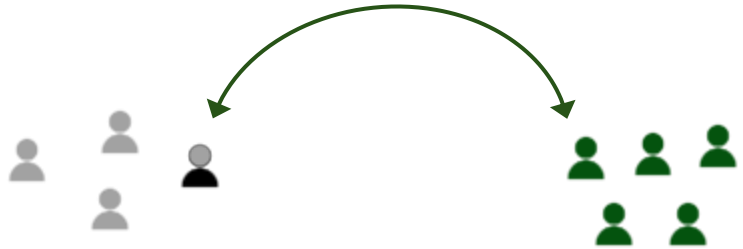
Part III: System

Odessa, a Decentralized Social Systems App

System

Odessa, a Decentralized Social Systems App





Part I

Community-centered norms

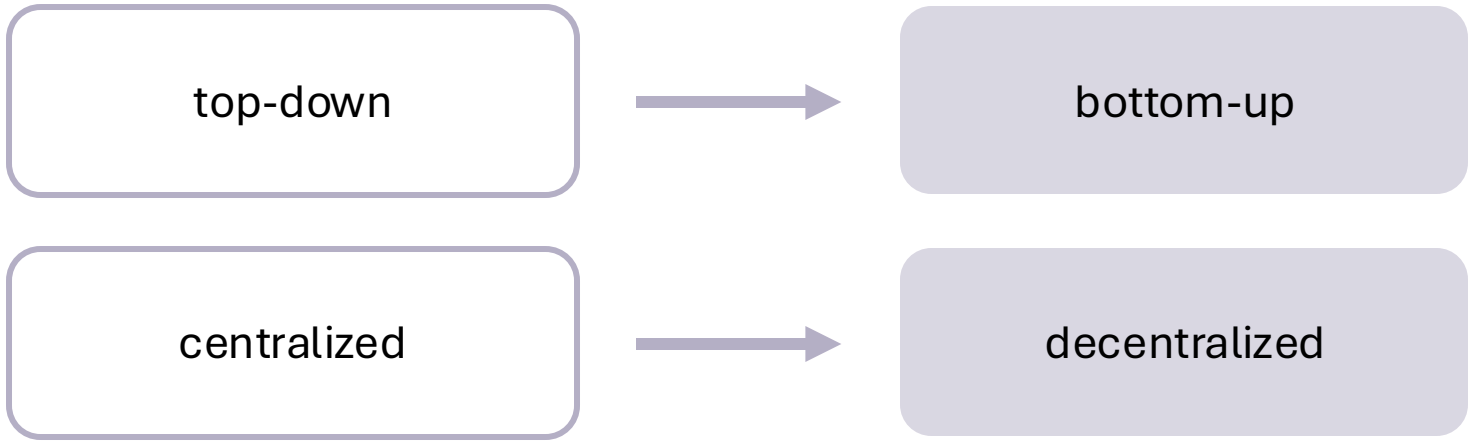
Part II

Explainable moderation

Understanding shared norms

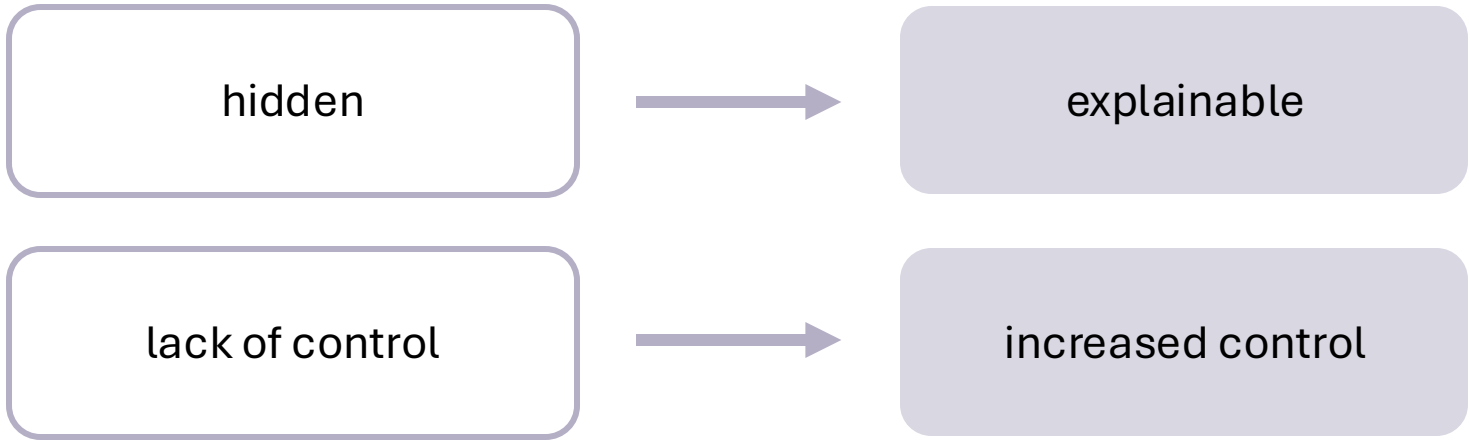
Moving towards decentralized principles

Odessa, a Decentralized Social Systems App



Focusing in algorithmic transparency

Odessa, a Decentralized Social Systems App



Building a more honest internet. Zuckerman. (2019)

System

Odessa, a Decentralized Social Systems App

Building a world where many worlds fit

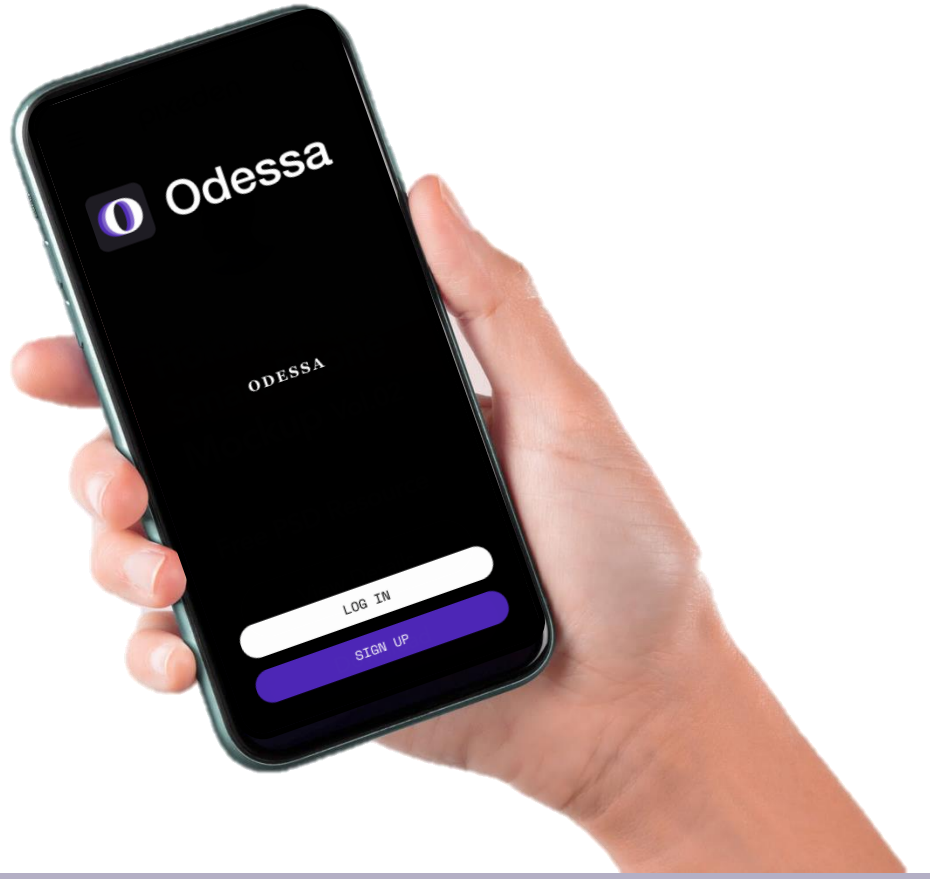
If there is to be an alternative,
we need to build it on purpose

Odessa: New research-focused social network!

A new system

Odessa allows researchers to empirically evaluate strategies for:

- ✓ defining community norms
- ✓ decentralized governance
- ✓ decentralized content moderation
- ✓ customized uplifting algorithms
- ✓ bridging communities



Odessa, a Decentralized Social Systems App!

- Open-source platform
- ~1.5 years of development
- ~1800 commits
- < 50k lines of code
- 100s users

O De S S A

Odessa, a **De**centralized **S**ocial **S**ystems **A**pp!

- Open-source platform
- ~1.5 years of development
- ~1800 commits
- < 50k lines of code
- 100s users

Collaborators

Lightweight sandbox

Easy to modify

Quick experimentation

Research questions

PhD Dissertation Scope

- ❑ How can Odessa facilitate community-centered governance?
- ❑ What design principles are needed to bridge communities with differing norms?
- ❑ What are challenges and opportunities in opening a bridging space?

Design considerations

MIT CCC–compatible suite

- **A.** Fora ~ Conversation-based network

AI–aided decentralized governance

- **B.** Community-centered norms
- **C.** Explainable content moderation
- **D.** Self-review process
- **E.** Uplifting content

Bridge space

- **F.** Shared prompts
- **G.** Shared governance mechanisms

Design considerations

MIT CCC–compatible suite

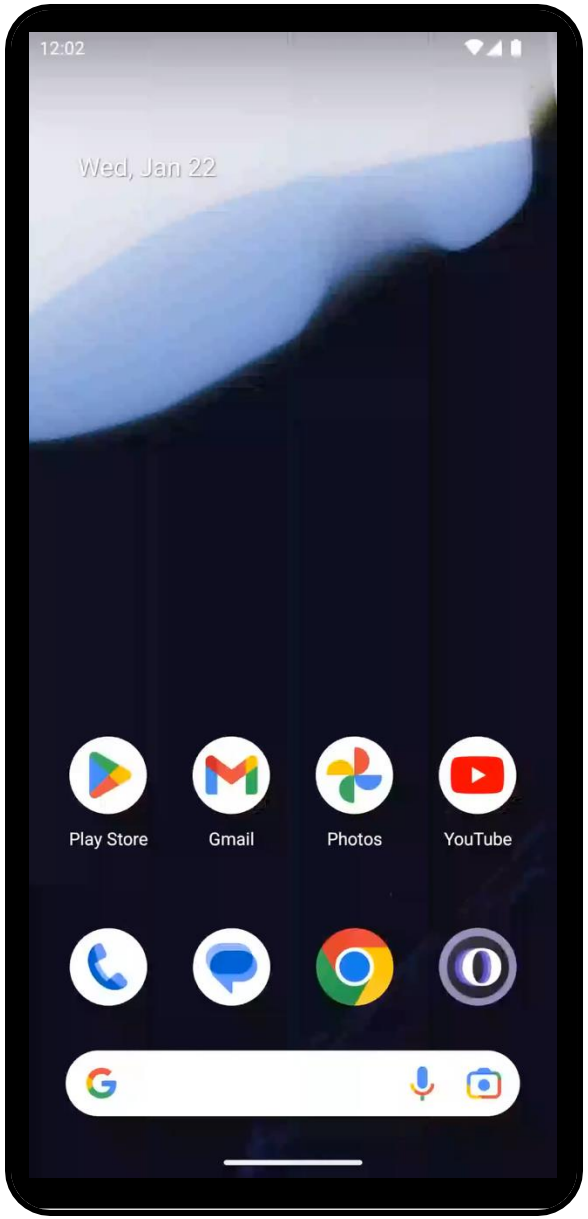
- **A.** Fora ~ Conversation-based network

AI–aided decentralized governance

- **B.** Community-centered norms
- **C.** Explainable content moderation
- **D.** Self-review process
- **E.** Uplifting content

Bridge space

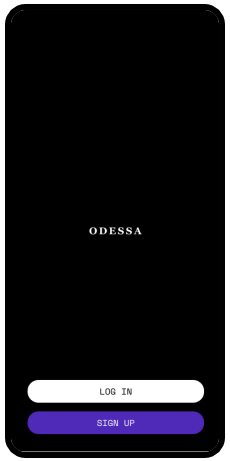
- **F.** Shared prompts
- **G.** Shared governance mechanisms



Contributions

1. Fully functional experimental social network

Design considerations



2. User study and empirical observations

Students

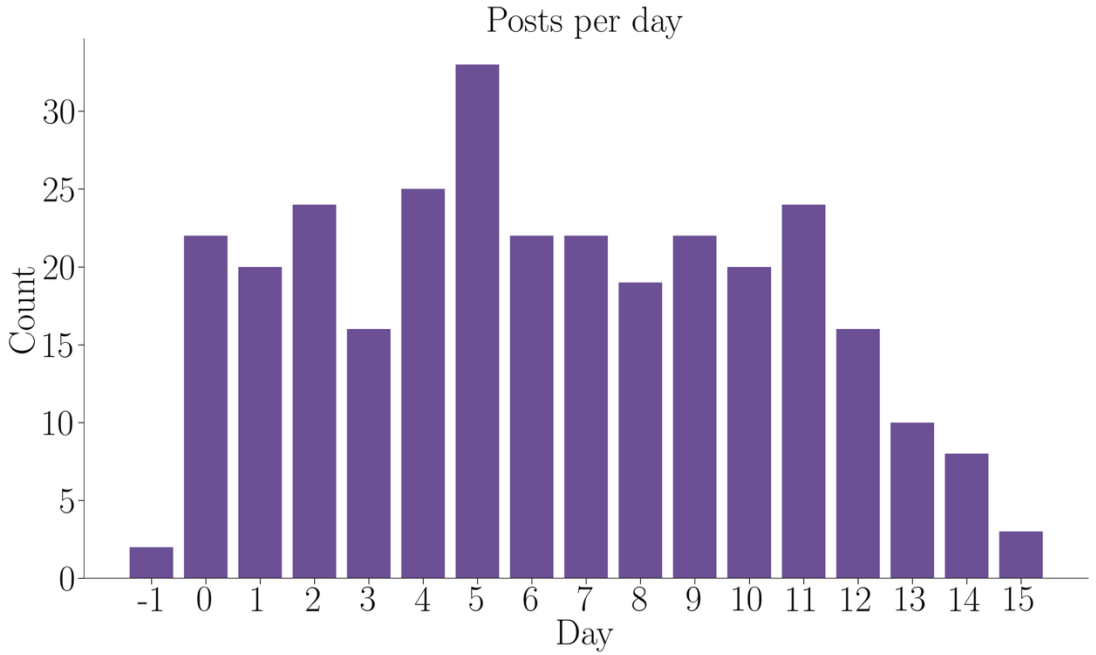
- ✓ MIT Grads
- ✓ MIT Undergrads
- ✓ Wellesley College

Human-subject experience

Empirical observations

- ✓ 34 participants
- ✓ 2 weeks (14 days)
- ✓ 3 communities
- ✓ 216 prompts
- ✓ 318 posts
- ✓ 522 chars per prompt

  ~30 seconds each



Design considerations

MIT CCC–compatible suite

- **A.** Fora ~ Conversation-based network

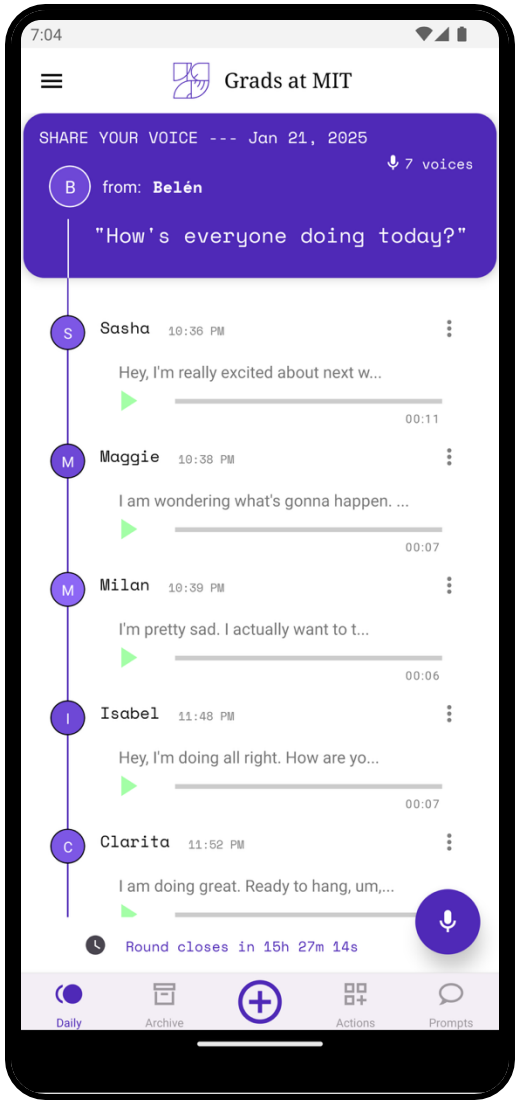
AI–aided decentralized governance

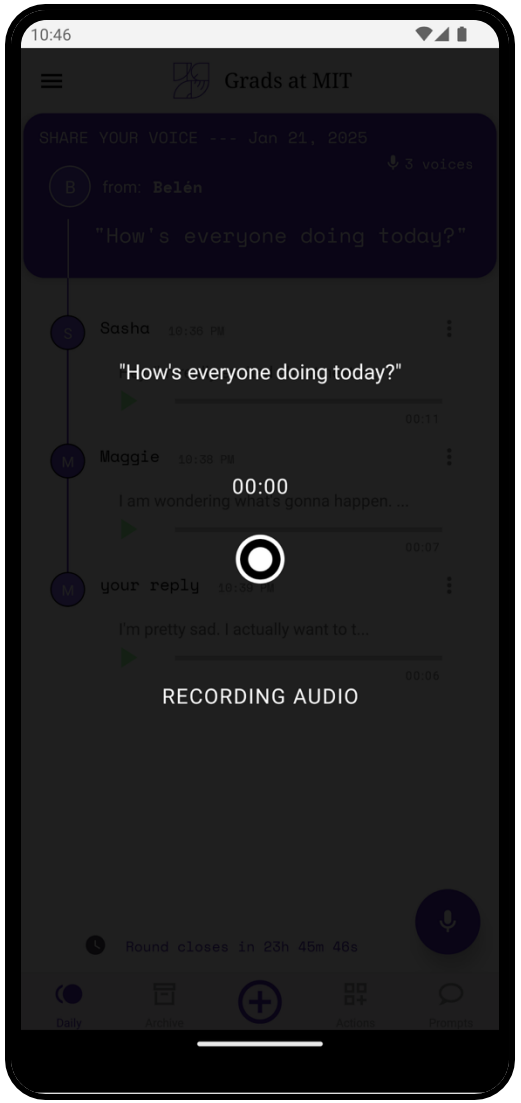
- **B.** Community-centered norms
- **C.** Explainable content moderation
- **D.** Self-review process
- **E.** Uplifting content

Bridge space

- **F.** Shared prompts
- **G.** Shared governance mechanisms

A. Fora ~ Conversation-based decentralized network



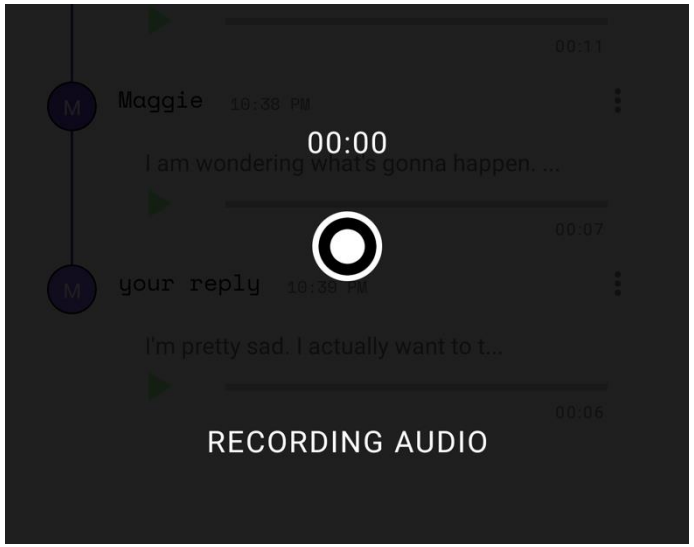


Compatible with

“it’s harder to degrade others when it has to be done aloud”

A. Fora ~ Conversation-based decentralized network

80% post-survey respondents, n=18



Contributions

- ✓ Voice-based social network, which transcribes voice notes immediately after posting and analyzes them with content moderation policies.

Implications

- ✓ Odessa requires users to “find time to record my voice”, making it a potentially less reactionary social network.

Design considerations

MIT CCC–compatible suite

- **A.** Fora ~ Conversation-based network

AI–aided decentralized governance

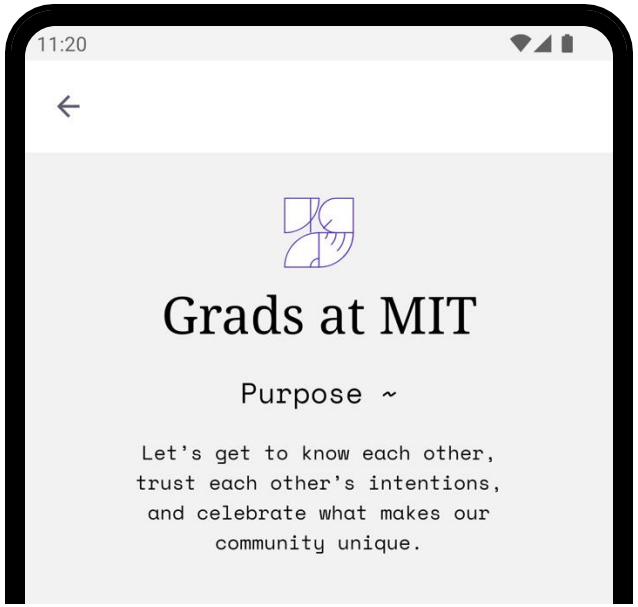
- **B.** Community-centered norms
- **C.** Explainable content moderation
- **D.** Self-review process
- **E.** Uplifting content

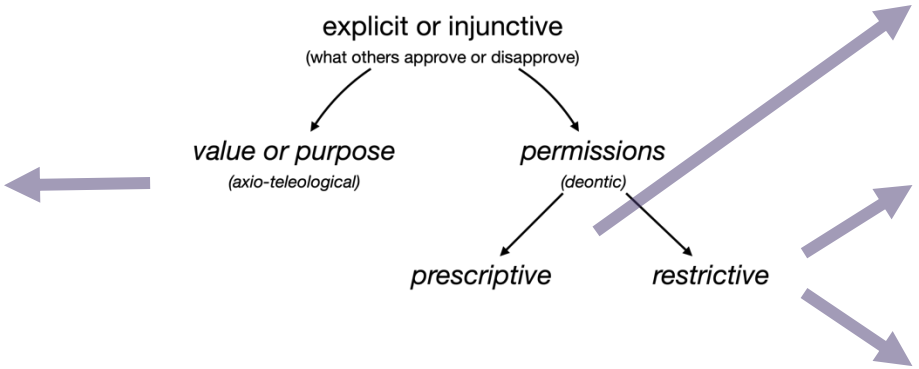
Bridge space

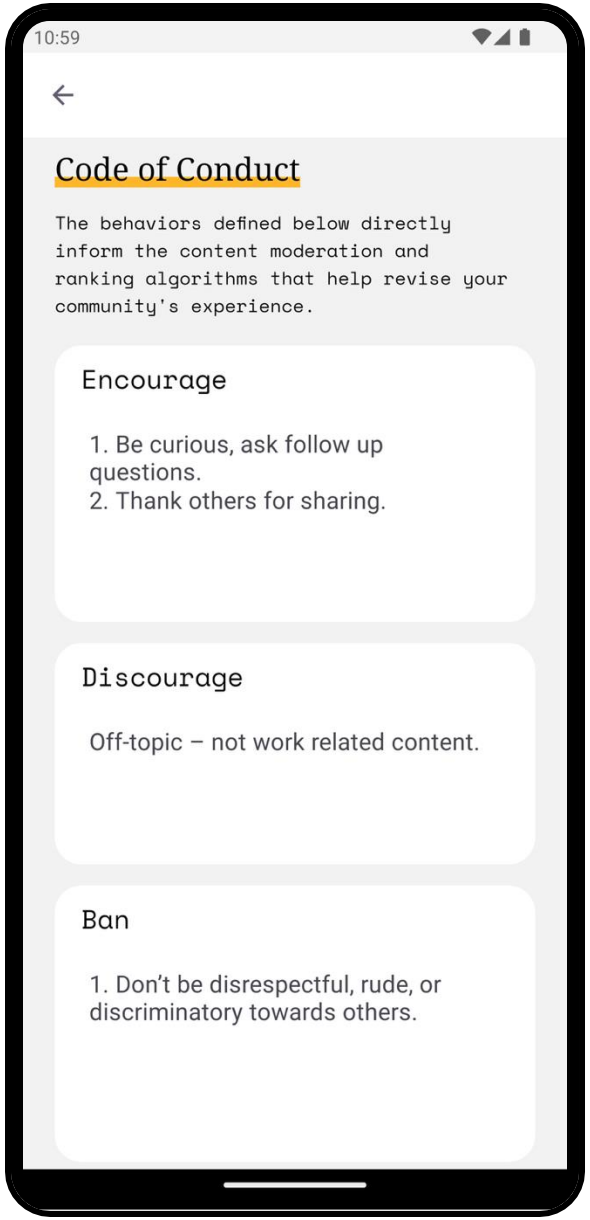
- **F.** Shared prompts
- **G.** Shared governance mechanisms

B. Community-centered norms

Speech norms and purpose set the tone







Opportunity: content policy design

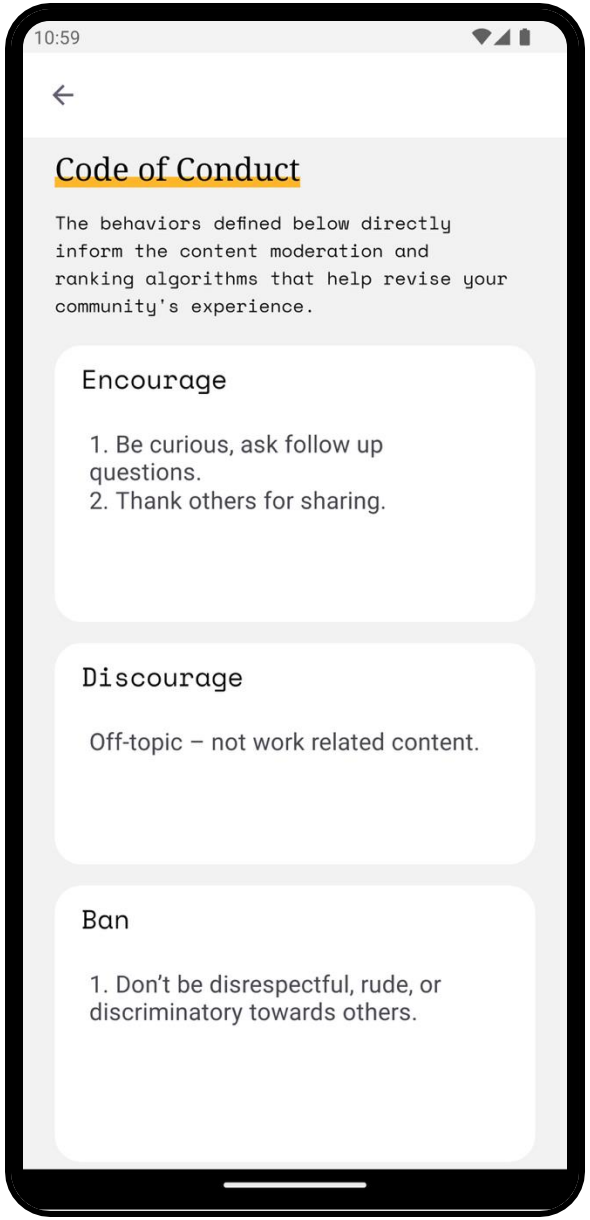
B. Moderation layer ~ Community-centered norms

Contributions

✓ Sandbox to test content policy development and user interaction with content moderation.

Implications

✓ Being explicit about the requirements to set up a community can set the tone.



Design considerations

MIT CCC–compatible suite

- **A.** Fora ~ Conversation-based network

AI–aided decentralized governance

- **B.** Community-centered norms
- **C.** Explainable content moderation
- **D.** Self-review process
- **E.** Uplifting content

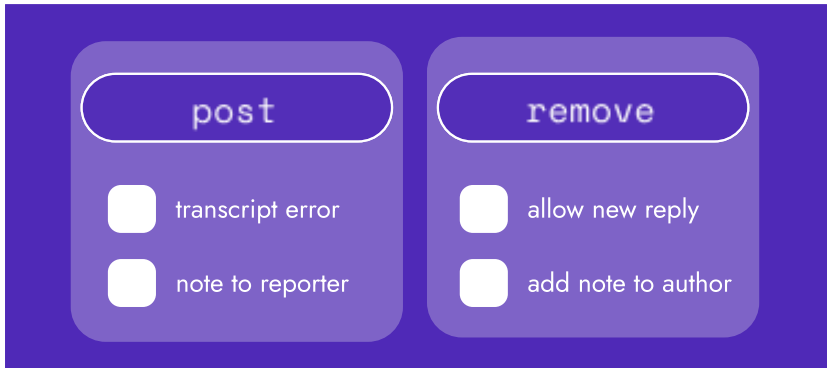
Bridge space

- **F.** Shared prompts
- **G.** Shared governance mechanisms

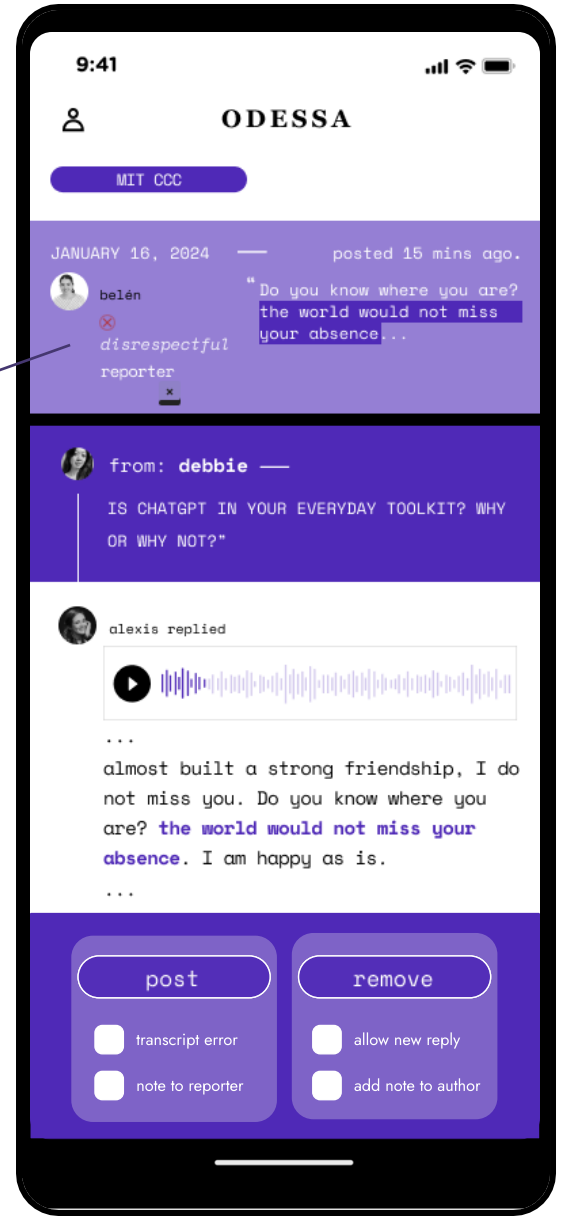
C. Explainable content moderation

not miss you. Do you know where you are? **the world would not miss your absence**. I am happy as is.
...

explainable moderation, **norm**:
don't be *disrespectful*.



moderator controllers

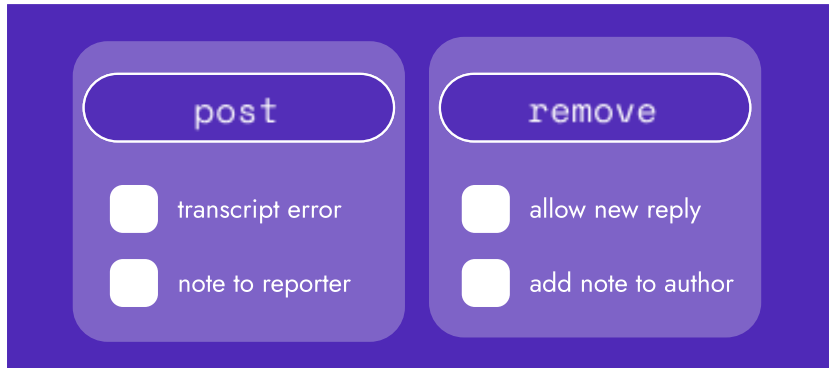


C. Moderation layer ~ Explainable content moderation

6 moderators

not miss you. Do you know where you are? **the world would not miss your absence**. I am happy as is.
...

explainable moderation, **norm**:
don't be *disrespectful*.



moderator controllers

Contributions

- ✓ *"highly efficient and easy to use"*
- ✓ Flexible to test different moderation algorithms and moderation scenarios

Implications

- ✓ *Governance-first approaches prioritize the experience of the moderation process, rather than leaving the moderation as a second thought.*

Current moderation challenges and opportunities

C. Moderation layer ~ Explainable content moderation

Challenges

- ❑ Enforcing norms differently
- ❑ Influence of faster moderators
- ❑ Conflict resolution

Implications

- ✓ Transparency
- ✓ Policy channel
- ✓ Forms of displaying moderated content
- ✓ Peer-review moderation strategies

Design considerations

MIT CCC–compatible suite

- **A.** Fora ~ Conversation-based network

AI–aided decentralized governance

- **B.** Community-centered norms
- **C.** Explainable content moderation
- **D.** Self-review process
- **E.** Uplifting content

Bridge space

- **F.** Shared prompts
- **G.** Shared governance mechanisms

D. Moderation layer ~ Self-review process

Post Guidance

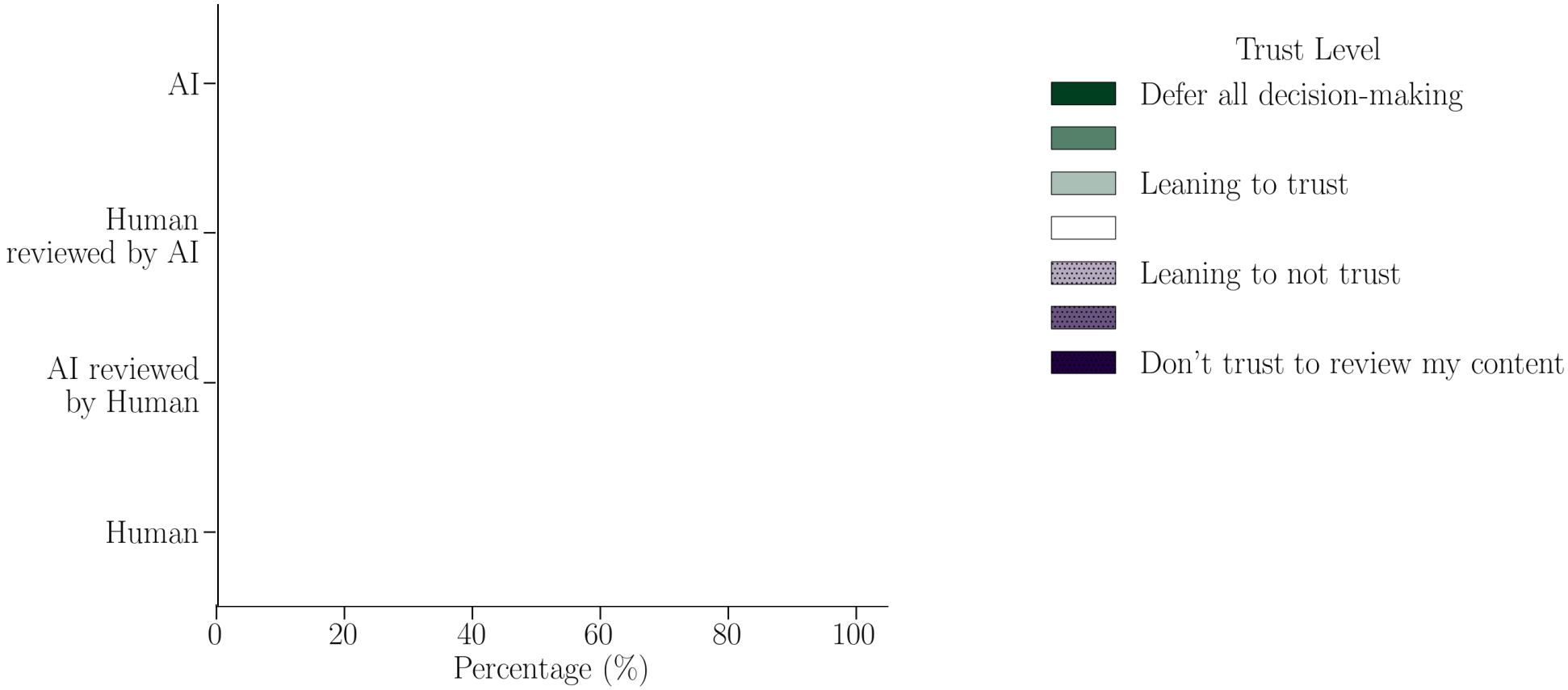
- ✓ Increased the number of “successful” contributions.
- ✓ Decreased moderation workload.

Post Guidance for Online Communities.
Manoel Ribeiro et al. (2025).

How much do you trust these to review your content?

D. Moderation layer ~ Self-review process

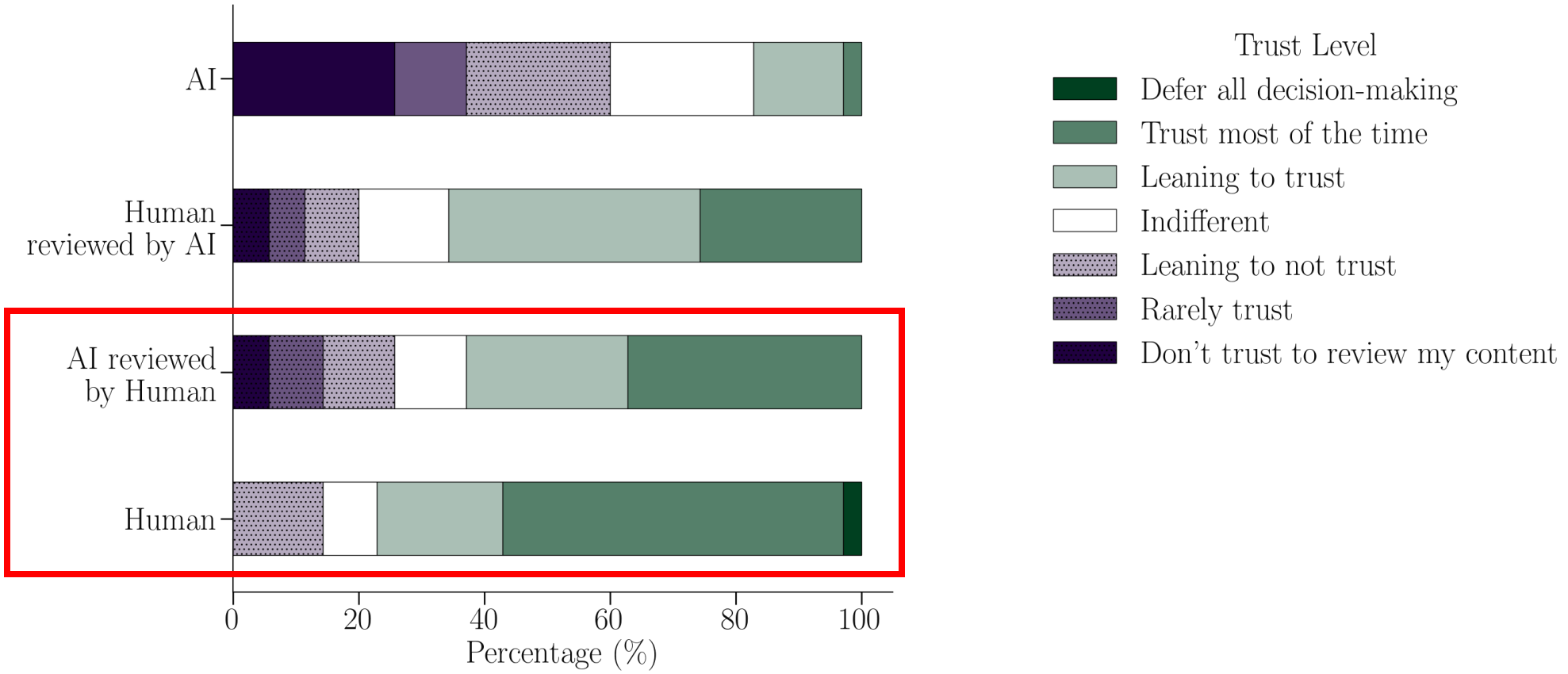
n = 34



How much do you trust these to review your content?

D. Moderation layer ~ Self-review process

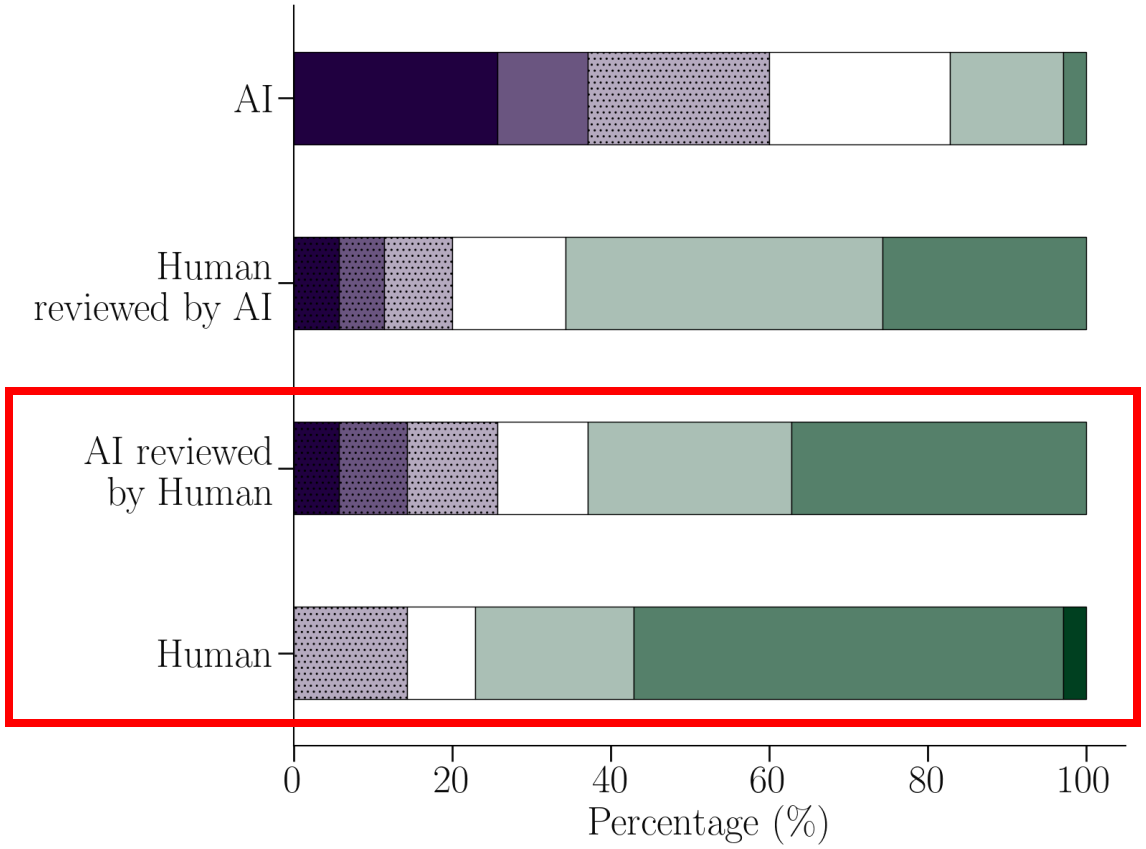
n = 34

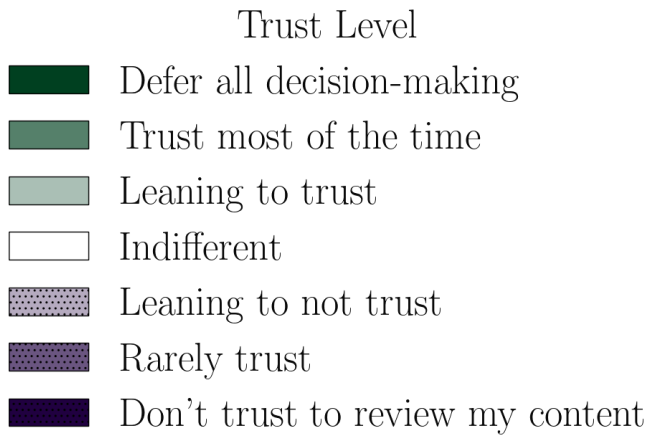


How much do you trust these to review your content?

D. Moderation layer ~ Self-review process

n = 34

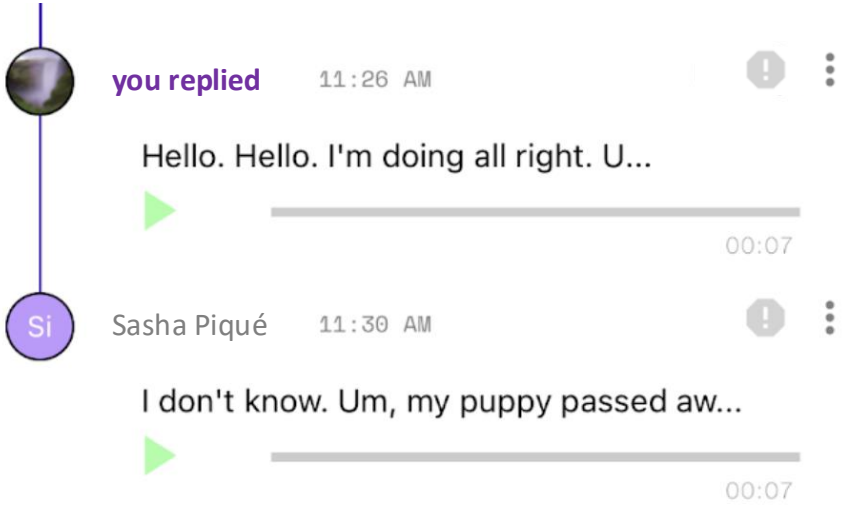




Implications

- ✓ Governance strategy depends on each community
- ✓ Explainable decision making from both humans and A.I.

D. Moderation layer ~ Self-review process



25%, found the process a burden

- X additional effort required
- X preference for other humans to handle

75%, proactively comfortable

- ✓ recognized its value for moderation
- ✓ fosters self-awareness
- ✓ promotes transparency and A.I. interaction

Design considerations

MIT CCC–compatible suite

- **A.** Fora ~ Conversation-based network

AI–aided decentralized governance

- **B.** Community-centered norms
- **C.** Explainable content moderation
- **D.** Self-review process
- **E.** Uplifting content

Bridge space

- **F.** Shared prompts
- **G.** Shared governance mechanisms

Bridging Communities

F. Shared prompts

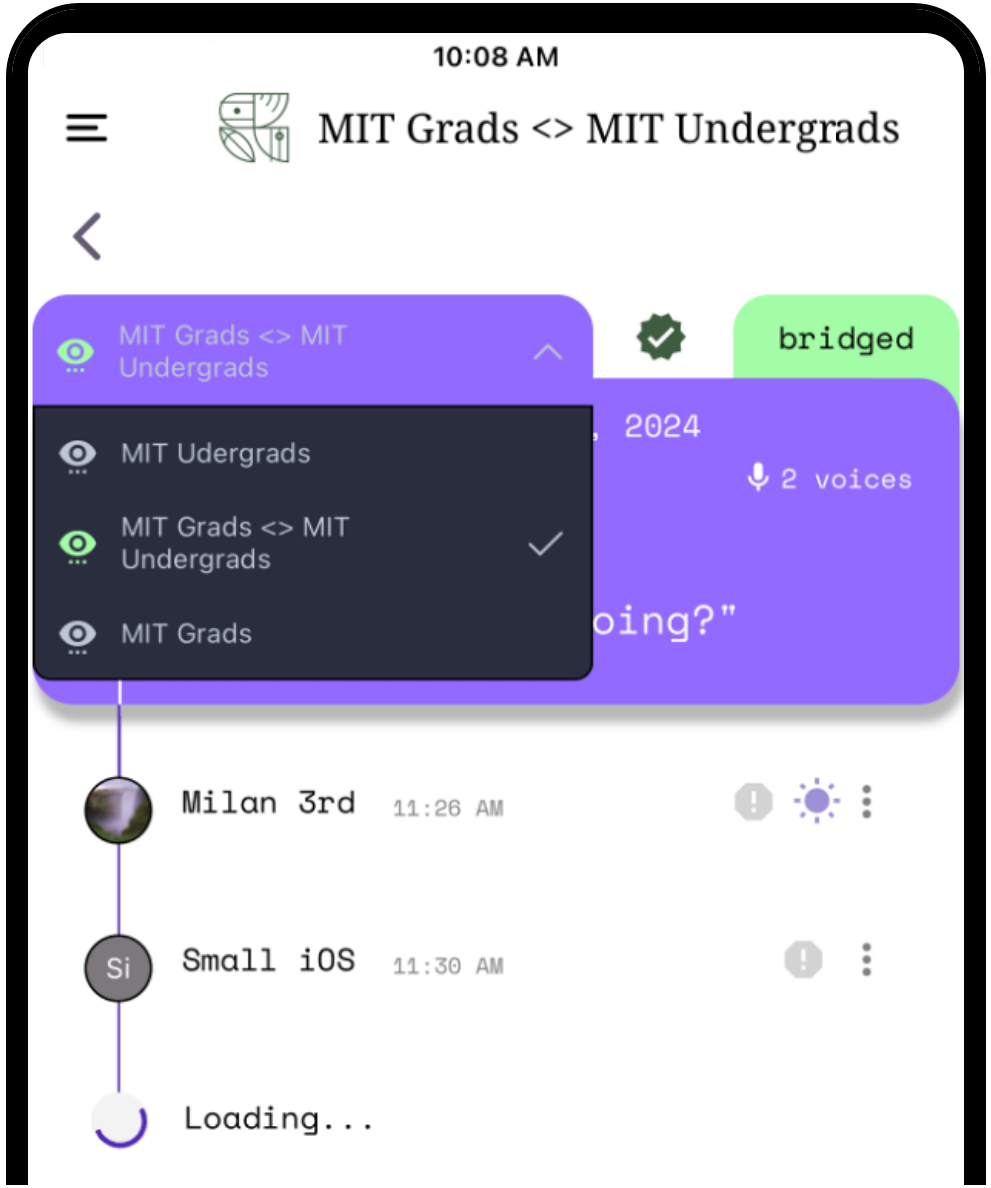
G. Shared governance

Implementation

- ✓ Filter per policy
- ✓ In shared prompt, report is reviewed by moderators from both communities

Opportunities

- ✓ Feedback opportunities through collecting reported comments in bridged prompt



Human-subject experience

Bridging communities

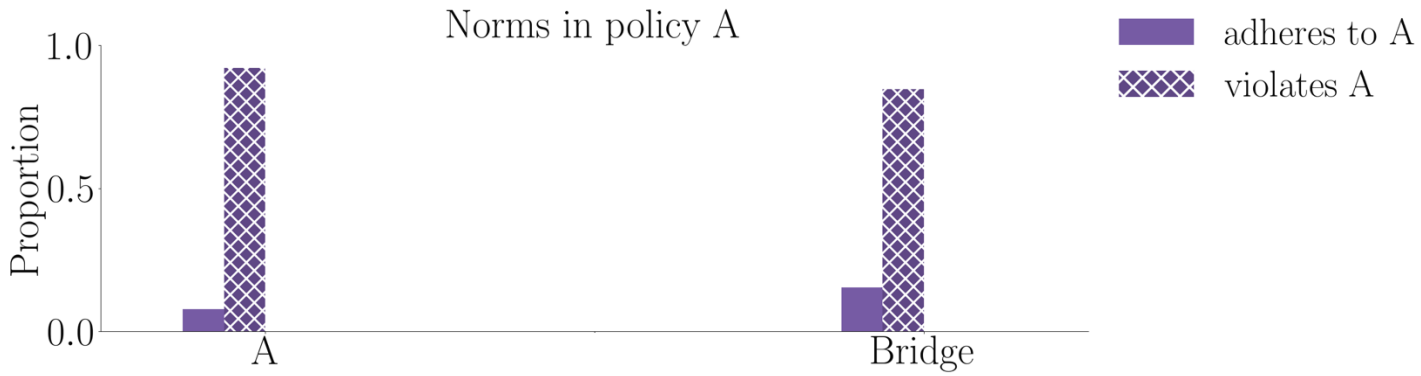
- ✓ 34 participants
- ✓ 2 weeks (14 days)
- ✓ **3 communities**
- ✓ 216 prompts
- ✓ 318 posts
- ✓ 522 chars per prompt

  ~30 seconds each

Community	Name	Norms
A	Takes about Life	1) Be respectful 2) Must introduce yourself whenever you post
B	Takes about Life in Boston	1) Be respectful 2) Don't share triggering stories
AB, Bridged	Hot Takes	<no restrictive norms>

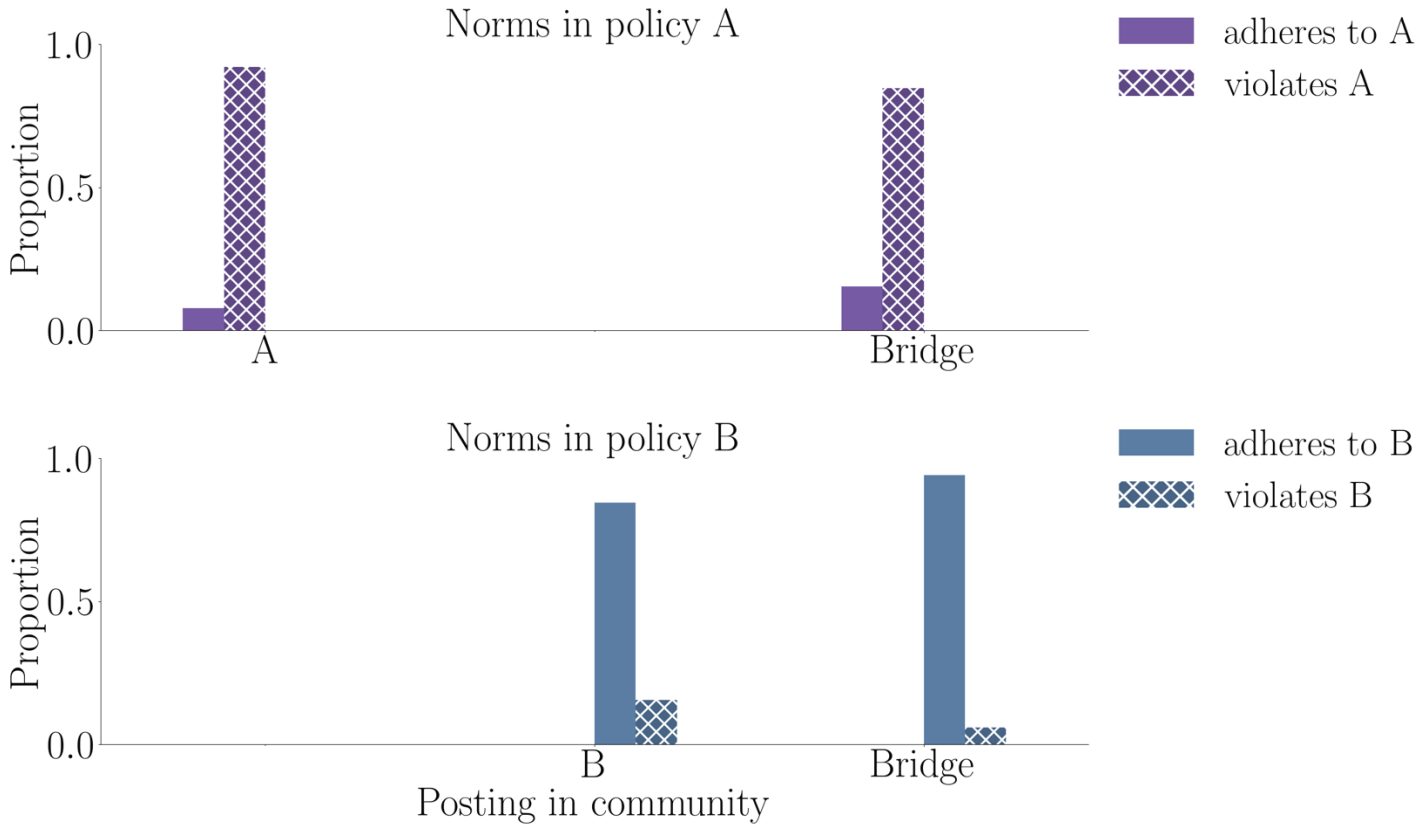
Users tend to carry their home community's behavior into the bridged space

Bridging communities



Users tend to carry their home community's behavior into the bridged space

Bridging communities



Observations

- Users are not used to reading the content policy often.
- Without bridge policy: No evidence to assume change in behavior without explicit visual cues or bridged policy.

Community and user–centered experience

Bridging communities

“I would prefer to have the communities separated”

“the convenience of having a public shared post makes it a lot easier than having to send a DM”

“I'm replying to my school friends so i don't think I need to keep my other friends in mind”

With Odessa,

- ✓ We have an opportunity to further imagine, prototype, and experiment with bridging designs and community-centered, or even person-centered governance and boundaries.

Contributions

- ✓ **New social network**: Fully functional.
- ✓ Layers of **distributed governance and content moderation**.
- ✓ Flexibility to incorporate any models to **uplift content**.
- ✓ **Empirical insights for bridging communities** based on speech norms.
- ✓ **A framework for hands-on learning about social media** algorithms and how we could make a change!

Contributions

Research Questions

RQ 1. How can we design opportunities for community-centered, explainable decentralized governance?

RQ 2. How can we design bridging opportunities for communities with differing norms and values?

- ✓ New social network: Fully functional.
- ✓ Layers of distributed governance and content moderation.
- ✓ Flexibility to incorporate any models to uplift content.
- ✓ Empirical insights for bridging communities based on speech norms.
- ✓ A framework for hands-on learning about social media algorithms and how we could make a change!

Towards Bridging and Governing Decentralized Communities

Part I: Data

Surfacing patterns from online communities

Part II: Tools

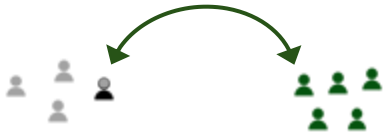
Self-governance and surfacing differences

Part III: System

Odessa, a Decentralized Social Systems App

Community-centered norms

Explainable moderation
Understanding shared norms



New social network
User study and design implications

Odessa: Decentralized Social Systems App

Towards Bridging and Governing Decentralized Communities

We need community–centered governance, moderation, and ranking.

Current approaches are not intentionally designed to bridge across divides.

Part I: Data

Surfacing patterns from online communities

✓ **Data**: We can use existent historical data to understand community values and norms at scale.

Part II: Tools

Self-governance and surfacing differences

✓ **Tools**: Current natural language processing tools can foster transparency and help uncover differences in perspectives across communities, helping us better understand each other.

Part III: System

Odessa, a Decentralized Social Systems App

✓ **System**: Bridging communities require an end-to-end approach, from data to infrastructure to user experience, with flexible governance structures. We offer a sandbox social media for this exploration!

Thanks to collaborators and funders

Deb K. Roy, Ph.D.

Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Rosalind W. Picard, Sc.D.

Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Jonathan L. Zittrain, J.D.

George Bemis Professor of International Law

Harvard University

MIT center for constructive communication

Project Liberty

BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

mit media lab

social machines

Thank you,

Dissertation Committee

Deb Roy, Ph.D.

Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Rosalind Picard, Sc.D.

Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Jonathan Zittrain, J.D.

George Bemis Professor of International Law

Harvard University