MICHIGAN ENGINEERING
SUMMER UNDERGRADUATE RESEARCH IN ENGINEERING
UNIVERSITY OF MICHIGAN

Value Alignment in Argumentation: What Predicts Shift in Perspective on r/ChangeMyView

Bhavesh Vuyyuru and Farnaz Jahanbakhsh
University of Michigan College of Engineering

INTRODUCTION

The ChangeMyView subreddit on Reddit serves as a platform where users present specific viewpoints and invite others to engage in discourse with the aim of potentially changing the original poster's (OP's) perspective. This study explores **value alignment**—the extent to which participants express similar values during a conversation—as a potential factor influencing these viewpoint shifts. Specifically, **we hypothesize that greater value alignment between the original poster (OP) and a commenter in the expressed arguments is positively associated with the probability that the OP will report a change in their view and award a delta**. As an initial step, we conducted a **pilot study** to begin to understand how value alignment between participants may relate to the likelihood of a shift in viewpoint.

METHODS

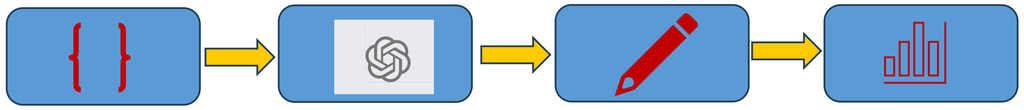
Pilot Study Setup:

Experimental Pipeline ran on subset of the dataset from *Cornell ChangeMyView Data*[1]

- Analyzed **23 posts** related to political and social issues
- **Goal:** Track values disclosed by OP and commentator as the conversation unfolds
- Values being tracked are from **Schwartz's 19 Basic Human Values**[2]
- Two runs performed:
  1. Threads containing OP and one commenter (230 total)
  2. Threads containing OP and one commenter **with** at least 3 comments (110 total)
- Each comment is rated one by one with entire preceding conversation given as context
- Store ratings for the 19 values as key value pairs in global value vectors for each user

Global Value Vector Update:

- Each user's first comment's rating is used to set the baseline global vectors
- For each succeeding comment's rating:
  1. **If rating = X** → no update
  2. **If rating is a 1 or 2** → take **maximum** of current value rating and global value rating
  3. **If rating = 0** → set global value rating to 0



Parse and filter data to isolate **two-user conversations.**

Each comment in each thread is rated for presence of values and their saliency by an **OpenAI API prompt**

Update **global value vectors** per user for each thread as comments are rated

Fit a **generalized linear mixed-effects model** to predict the probability of a delta being awarded based on value alignment, while accounting for variation across posts.

Comment Scoring System

Each value is given a rating:

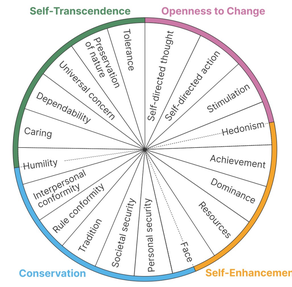
- **X**: Value not expressed
- **0**: Value opposed
- **1**: Value weakly supported
- **2**: Value strongly supported

Value Misalignment Measurement: Mean Absolute Error

MAE between two user vectors reflects the average difference in value expression between users → measures misalignment

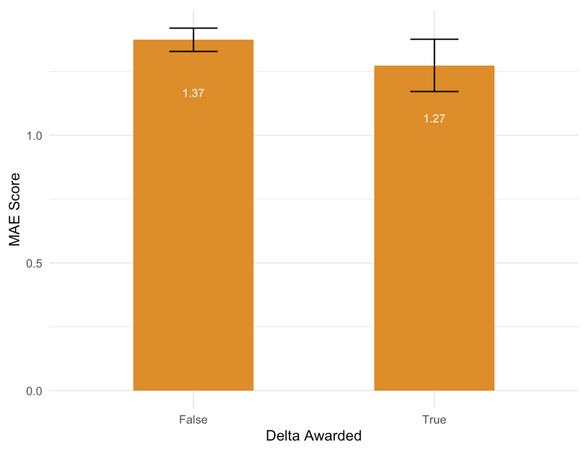
- Higher scores indicate greater divergence in expressed values
- A score of **0** indicates **perfect alignment**
- Only **shared indices with non-X values** are included
- Penalty of 2 (max difference) applied to shared indices with one X and non X value

We classify the presence and saliency of values by drawing from Schwartz's Theory of 19 Basic Human Values[2]



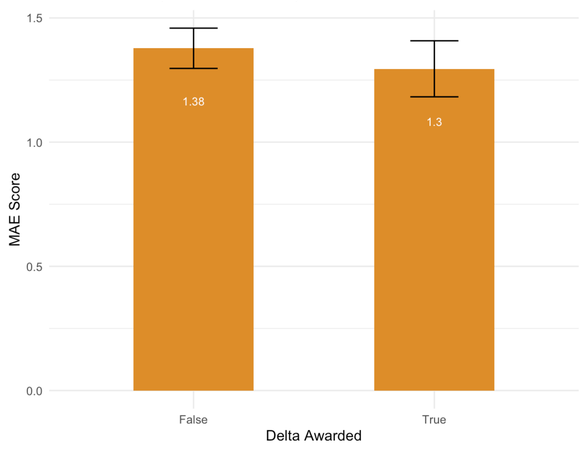
RESULTS

MAE Score Means and 95% Confidence Intervals by Delta Awarded (All Threads)



- This plot represents data containing all threads (n=230 FALSE=190 TRUE=40)
- Average Misalignment Score was 0.1 lower in threads where a delta was awarded

MAE Score Means and 95% Confidence Intervals by Delta Awarded (Filtered Threads)



- This plot represents data containing threads ≥ 3 comments (n=110 FALSE=70 TRUE=40)
- Average score was 0.08 lower in threads where a delta was awarded

Summary of Fixed Effects from Generalized Linear Mixed-Effects Model

Type	MAE Score	Odds Ratio	P-value
All Threads	-1.03	0.358	0.069
All Threads ≥ 3 comments	-0.813	0.444	0.198

- A 1-unit increase in MAE Score is associated with a **decrease** in the log-odds of delta being awarded by about 1.03 and 0.813 respectively.
- A 1 unit increase in score is associated with ((1-Odds Ratio) *100) percent decrease in a delta being awarded

CONCLUSION

Analysis of Pilot Study Results:

- **Lower mean** MAE score in threads where delta was awarded indicates **higher alignment**
- **Higher alignment** was associated with **increased likelihood** of delta being awarded

Future Directions:

- Expand study to broader data to determine statistical significance of differences observed
- Operationalize a framework for quantifying alignment among three or more users.

REFERENCES

[1] Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, Lillian Lee (WWW'2016).

[2] Kolluri, Akaash, Renn Su, Farnaz Jahanbakhsh, Dora Zhao, Tiziano Piccardi, and Michael S. Bernstein. "Alexandria: A Library of Pluralistic Values for Realtime Re-Ranking of Social Media Feeds." *arXiv preprint arXiv:2505.10839* (2025).