

박사학위논문
Ph.D. Dissertation

해석적 스캐폴딩을 통한 사회적 이슈에 대한 정보
기반 의견 형성 지원

Facilitating Informed Opinion Formation on Social Issues
through Interpretive Scaffolds

2026

김현우 (金玄宇 Kim, Hyunwoo)

한국과학기술원

Korea Advanced Institute of Science and Technology

박 사 학 위 논 문

해석적 스캐폴딩을 통한 사회적 이슈에 대한 정보
기반 의견 형성 지원

2026

김 현 우

한 국 과 학 기 술 원

전산학부

해석적 스캐폴딩을 통한 사회적 이슈에 대한 정보 기반 의견 형성 지원

김 현 우

위 논문은 한국과학기술원 박사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2026년 5월 29일

심사위원장 김 주 호 (인)

심 사 위 원 장 정 우 (인)

심 사 위 원 Joseph Seering (인)

심 사 위 원 Amy X. Zhang (인)

심 사 위 원 Brian McInnis (인)

Facilitating Informed Opinion Formation on Social Issues through Interpretive Scaffolds

Hyunwoo Kim

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Daejeon, Korea
June 17, 2026

Approved by

Juho Kim
Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics¹.

¹ Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

DCS

김현우. 해석적 스캐폴딩을 통한 사회적 이슈에 대한 정보 기반 의견 형성
지원. 전산학부 . 2026년. 149+xi 쪽. 지도교수: 김주호. (영문 논문)

Hyunwoo Kim. Facilitating Informed Opinion Formation on Social Issues
through Interpretive Scaffolds. School of Computing . 2026. 149+xi pages.

Advisor: Juho Kim. (Text in English)

초 록

사회적 이슈에 대한 정보 기반 의견 형성은 정보 접근만으로 이루어지지 않는다. 개인은 주장을 그 너머의 맥락에서 해석하고, 전제된 가정을 인식하며, 경쟁하는 관점에 비추어 근거를 따져야 한다. 그러나 사람들은 노력이 드는 추론을 회피하고, 동기가 있어도 필요한 맥락을 스스로 떠올리지 못하며, 오늘날의 정보 환경은 그 맥락을 오히려 분절시킨다. 본 학위논문은 추론자와 정보가 만나는 순간에 개입하여 정보를 해석·검토·통합하는 대안적 방식을 제시하는 설계 개념인 해석적 스캐폴드(interpretive scaffold)를 제안한다. 해석적 스캐폴드는 외부 맥락으로 정보를 확장하는 보강형과 정보에 내재한 측면을 드러내는 재구조화형으로 작동한다. 이 개념은 네 시스템으로 구현된다. StarryThoughts는 의견의 인구통계적·이념적 지형을, DataDive는 통계적 주장 검토용 외부 데이터를, FrameIt은 의견에 내재한 도덕적 기반을 드러내고, DecideTogether는 운영자가 의도를 규칙으로 표현하도록 돕는다. 사용자 연구와 기술적 평가를 통해, 해석적 스캐폴드가 추론에 유입되는 정보와 그 처리 방식을 변화시켜 다양한 의견에 대한 존중적 참여, 정보의 비판적 소비, 자신의 정보 기반 입장 표현을 촉진함을 보인다.

핵심 낱말 해석적 스캐폴드, 정보 기반 의견 형성, 정보 해석, 맥락적 추론, 센스메이킹, 비판적 사고, 숙의, 논쟁적 사회 이슈, 인간 중심 AI, 소셜 컴퓨팅

Abstract

Forming informed opinions on social issues requires more than access to information. Individuals must interpret claims in light of more than the text itself, recognizing the assumptions on which an argument rests and weighing evidence against competing perspectives. The Internet offers unprecedented access to information, yet research in political psychology and cognitive science shows that people struggle to do this work: they are often unmotivated to reason effortfully and fall back on heuristics and motivated reasoning, they fail to generate relevant context on their own even when motivated, and contemporary information environments fragment context rather than supply it.

This dissertation proposes interpretive scaffolds, a design construct for supporting individual reasoning at the moment of encounter between a reasoner and a piece of information by surfacing alternative ways to interpret, examine, and integrate it. Interpretive scaffolds operate in two modes: enriching scaffolds broaden the information under consideration with external context, and restructuring scaffolds surface latent facets within it. I instantiate this construct through four interactive systems. StarryThoughts surfaces the demographic and ideological landscape of perspectives on an issue. DataDive provides relevant external data to support critical assessment of statistical claims. FrameIt makes the moral foundations embedded in online opinions visible. DecideTogether supports articulation in online community moderation, helping moderators translate community-specific intents into concrete rules by analyzing community context and system-generated interpretations of those intents.

Across these systems, user studies and technical evaluations show that interpretive scaffolds change what information enters reasoning and how people engage with it. Together, this thesis demonstrates that interpretive scaffolds, by enriching information with broader context and surfacing its latent structure,

foster respectful engagement with diverse opinions, critical consumption of information, and articulation of one's own informed position on contested social issues.

Keywords interpretive scaffolds, informed opinion formation, information interpretation, contextual reasoning, sensemaking, critical thinking, deliberation, contested social issues, human-centered AI, social computing

Contents

Contents	i
List of Tables	vii
List of Figures	ix
Chapter 1. Introduction	1
1.1 Informed Opinions	1
1.2 Challenges in Forming Informed Opinions	1
1.3 Interpretive Scaffolds as Facilitators of Individual Reasoning . .	2
1.4 Instantiations of Interpretive Scaffolds	3
1.5 Contributions	4
1.6 Thesis Overview	4
Chapter 2. Background	6
2.1 Informed Opinion and its Formation	6
2.1.1 Informed Opinion in Democratic Theory	6
2.1.2 The Individual Work of Deliberation	6
2.1.3 How Individuals Form Informed Opinions	7
2.2 Challenges with Forming Informed Opinion Formation	7
2.3 Interactive Systems for Supporting Information Engagement .	8
2.3.1 Surfacing the broader information space	8
2.3.2 Making Latent Structure Visible	9
2.3.3 Combining Enrichment and Restructuring	10
Chapter 3. StarryThoughts: Facilitating Diverse Opinion Exploration on Social Issues	11
3.1 Motivation & Contributions	11
3.2 Initial Design & Iteration	12
3.2.1 L1: Inspecting fine-grained demographic groups	14
3.2.2 L2: Leveraging users' prior knowledge for engagement .	14
3.2.3 L3: Active guidance for exploring diverse opinions . . .	14
3.3 System	14
3.3.1 System usage scenario	17
3.3.2 Recommendation algorithm (L3)	18
3.3.3 Implementation details	18
3.4 Evaluation	19

3.4.1	Participants	19
3.4.2	Conditions	19
3.4.3	Procedure	20
3.4.4	Measures	23
3.5	Results	24
3.5.1	Strategies and patterns of exploration	25
3.5.2	Effect of using StarryThoughts for opinion strength and quality	27
3.5.3	Effect of StarryThoughts on evaluation of other's opinions	29
3.6	Discussion	32
3.6.1	Role of identity	32
3.6.2	Effect of visualizing opinions	33
3.6.3	Value of recommendation	34
3.6.4	Generalization to other issues and cultures	35
3.6.5	Limitations	36
3.6.6	Future Work	36
3.7	Conclusion	36
Chapter 4. DataDive: Supporting Readers' Contextualization of Statistical Statements with Data Exploration		37
4.1	Motivation & Contributions	37
4.2	Formative Study	39
4.2.1	Study Setup	39
4.2.2	Design Goals	40
4.3	User Interaction	41
4.3.1	Step 1. Initiating Context Exploration	41
4.3.2	Step 2. Browsing Recommended Contexts	41
4.3.3	Step 3. Investigating and Exploring Data	42
4.4	Technical Pipeline	43
4.4.1	The Context Database	43
4.4.2	Stage 1. Generation of Ranked Candidate Contexts . .	44
4.4.3	Stage 2. Extraction of Relevant Data	45
4.4.4	Statistical Check-worthiness	46
4.5	Pipeline Evaluation	47
4.5.1	Materials	47
4.5.2	Participants	47
4.5.3	Procedure	47
4.5.4	Results	47

4.6	User Evaluation	49
4.6.1	Participants	49
4.6.2	Conditions	49
4.6.3	Procedure	50
4.7	User Evaluation Results	51
4.7.1	Assessment of RQ1	51
4.7.2	Assessment of RQ2	54
4.7.3	Assessment of RQ3	56
4.8	Discussion	57
4.8.1	Incorporation of Contextual Information Beyond Structured Data	57
4.8.2	Considerations for Factually Reliable LLM-Powered System	58
4.8.3	Considerations for an Organically Growing Context DB	58
4.8.4	Limitations and Future Work	59
4.9	Conclusion	59

Chapter 5.	Framelt: Supporting Respectful Online Discussions through Moral Foundations	60
5.1	Motivation & Contributions	60
5.2	Formative Study	63
5.2.1	Design Goals	63
5.3	FrameIt	64
5.3.1	Use Case Scenario	64
5.3.2	System Implementation	67
5.4	Technical Evaluation	69
5.4.1	Methods	69
5.4.2	Results	70
5.5	User Study 1: Comparative Lab Study	71
5.5.1	Participants	71
5.5.2	Conditions	72
5.5.3	Procedure	72
5.5.4	User Study Measures	73
5.6	User Study 1 Results	73
5.7	User Study 2: Deployment-based User Study	75
5.7.1	Participants	78
5.7.2	Materials	78
5.7.3	Procedure	79

5.8	User Study 2 Results	79
5.8.1	Change in Discussion Atmosphere	80
5.8.2	Scaffolding Respectful Communication	80
5.8.3	Tensions	82
5.9	Discussion	83
5.9.1	Theory of Change	83
5.9.2	Agreeing to Disagree: Respect at the Cost of Contestation	84
5.9.3	Conceptual Generalization Beyond Moral Foundations .	84
5.9.4	Ethical Implications	85
5.10	Limitations and Future Work	86
5.11	Conclusion	86
Chapter 6.	DecideTogether: Supporting Automoderation with Subjective Criteria through Human-AI Collaborative Calibration	88
6.1	Motivation & Contributions	88
6.2	Design Goals	90
6.2.1	DG1: Surface Latent Criteria in Community Context .	90
6.2.2	DG2: Make Automoderation Results Usable for Intent Articulation	90
6.2.3	DG3: Keep Rule Authoring and Revision Legible With- out Technical Expertise	91
6.2.4	DG4: Make Moderation Decisions Transparent and Trace- able	91
6.3	System	91
6.3.1	Key Idea: Aligning Rule Text and Automoderator Logic	91
6.3.2	User Interaction	92
6.3.3	Technical Pipeline	99
6.4	Technical Evaluation	101
6.4.1	Q1: Quality of Automoderator Logic	103
6.4.2	Q2 and Q3: Moderation Performance	104
6.5	User Study	106
6.5.1	Participants	106
6.5.2	Scenarios	106
6.5.3	Procedure	107
6.6	Results	108
6.6.1	RQ1: How does DecideTogether support community moderators in articulating and refining moderation in- tent?	108

6.6.2	RQ2: How does DecideTogether affect moderators' understanding, trust, and control over automoderation rules?	110
6.7	Discussion	111
6.7.1	Intent Articulation through Reviewing Real-World Cases	111
6.7.2	Decomposition as a Substrate for Inspection	112
6.7.3	Trust and Human Authority in Subjective Automoderation	113
6.8	Limitation and Future Work	113
6.9	Conclusion	114
Chapter 7.	Discussion	115
7.1	Synthesis of Empirical Support for Informed Opinion Formation	115
7.2	Design Reflections on Interpretive Scaffolds	116
7.2.1	Scaffolding Types: Should Address the Interpretive Bottleneck	116
7.2.2	Level of Interpretive Mediation: Match System Effort to User Capability	117
7.2.3	Interaction Design: from Lightweight Nudge to Deep Guidance	118
7.3	Potential Biases and Risks from Interpretive Scaffolds	118
7.3.1	Framing by Design	119
7.3.2	Bias from the AI Pipeline	119
7.3.3	Legitimation of Bad-Faith Content	119
7.3.4	Scaffolds and Interpretive Skill	120
7.4	Interpretive Scaffolds for the Social Process of Deliberation . .	120
7.4.1	Respectful Engagement Across Disagreement	121
7.4.2	Mutual Understanding	121
7.4.3	Inclusive Participation	121
7.4.4	A Shared Basis for Collective Judgment	122
7.4.5	Summary	122
7.5	Toward User-Driven Interpretive Scaffolds	122
7.6	Limitations	123
Chapter 8.	Conclusion	124
8.1	Summary of Contributions	124
8.2	Future Directions	124
8.2.1	Longitudinal Evaluation of Interpretive Scaffolds	124

8.2.2	Interpretive Scaffolds for Agent-Mediated Information Consumption	125
8.2.3	Multimodal Information	125
	Bibliography	126
	Acknowledgments	148

List of Tables

3.1	Perceived importance and prior knowledge level of the issues used in the study. Results suggest that the participants considered the Retirement issue as more important but knew less about it. (Wilcoxon signed-rank test, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$) . . .	24
3.2	Characteristics of opinions read by the participants from the first round of evaluation. The numbers suggest that people were more interested in the opinions of their own groups.	25
3.3	Argument repertoire of the participants before and after using the system. The participants were divided into three groups according to their prior opinion on the issue: supporters, neutrals, and dissenters. The table shows the number of reasons supporting and disagreeing with participants' own stance. For the neutral group, we reported the number of reasons that supported the issue in the column "Reasons for own opinion". The numbers show that the argument repertoire of the participants increased after using the system. (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$)	27
3.4	Distribution of opinions on the retirement issue from the participants before and after using the system. While more than half of the participants kept their opinion, it could be observed that some of the participants with strong prior opinions became moderate after using the system.	28
3.5	Distribution of opinions on the Game Addiction issue from the participants before and after using StarryThoughts. The results show that some of the participants with strong prior opinions became moderate after using the system.	28
3.6	Comparison between the design of StarryThoughts and existing systems.	33
4.1	The rules of DataDive for generating visualization.	42
4.2	Criteria for generating candidate contexts.	45
4.3	The average behavior counts to explore external information. The difference was significant for the Fertility Rate article and for the total.	51
4.4	The number of recalled questions per each article and condition. We did not observe statistically significant differences.	52
4.5	Participants' self-rating on exploring external information. There was no statistical significance between the DataDive condition and the baseline condition.	52
4.6	The number of recalled facts for each article and condition. The values are reported as DataDive vs. Baseline. We did not observe statistically significant differences between conditions, except for the number of recalled facts out of the text in the Carbon Emission article and in total.	54
4.7	Participants' self-rating on their gain after the experiment. There was no statistical significance between DataDive and the baseline condition.	55
4.8	Participants' task load for each condition with the NASA-TLX questionnaire. There was no statistical significance between DataDive and the baseline condition.	56
5.1	The number of occurrences per moral foundation and the likelihood in the ground truth. .	69

5.2	The precision, recall, and F1 score per each moral foundation. Our evaluation results show that the our LLM prompts showed performance comparable to the SOTA model. .	70
5.3	Demographic profiles of the user study participants and their prior opinions on the two issues used in the experiment. The prior opinions were measured in a 7-point Likert scale (1: Strongly Disagree, 7: Strongly Agree).	71
5.4	Pairwise comparison of self-evaluated discussion experience between FrameIt and baseline conditions. Questions with statistical significance were bolded.	74
5.5	Comparison of self-evaluated discussion experience between FrameIt and baseline conditions on Single-use Plastic Ban issue. We did not observe any statistically significant difference.	74
5.6	Comparison of self-evaluated discussion experience between FrameIt and baseline conditions on Cancel Culture. Questions with statistical significance were bolded.	75
5.7	Pre-post difference of self-evaluated receptiveness of the participants on Single-use Plastic Ban issue. While there was no significant difference in the baseline condition, participants in FrameIt condition had positive changes in the perceived informativeness and curiosity to different opinions.	76
5.8	Pre-post difference of self-evaluated receptiveness of the participants on Cancel Culture issue. While there was no significant difference in the baseline condition, participants in FrameIt condition had positive changes in the willingness to have conversation with people with opposing views.	77
5.9	Demographic profiles of the field study participants and their prior opinions on the three issues used in the experiment. The prior opinions were measured in 7-point Likert scale (1: Strongly Disagree, 7: Strongly Agree).	78
6.1	The set of subreddits we used for the technical evaluation, and the number of rules for each subreddit	102
6.2	Functional accuracy of the DecideTogether-generated automoderation logic. (15 subreddits, $n=2146$). Best non-trivial value per column in bold	104
6.3	Previous moderation experiences of the user study participants	106
7.1	Design space for interpretive scaffolds across the four systems.	116

List of Figures

3.1	The initial version of the system. The initial version of the system presents the variability and diversity of opinions by plotting them as dots in a 2-D space.	13
3.2	Prediction screen for the user. Before seeing the visualization of the opinions, the user is first asked to share their prediction on the groups who would likely support or object to the issue. The user can quickly verify their prediction by seeing actual opinions in the next step.	15
3.3	An overview of StarryThoughts. Each opinion is displayed as a dot in a two-dimensional space. The user's opinion is highlighted in red, and recommended opinions are highlighted in yellow, blue, or green according to the recommendation type. (A) With the filters, the user can specify demographic groups of interest. (B) The user can quickly test their predictions on who would support or oppose the issue with the quick filters. (C) By clicking or hovering on one of the dots, the user can read the opinion along with the demographic profile of the author.	15
3.4	Demographic filter of StarryThoughts. Each available option is presented as a dropdown menu at the upper left corner of the page. The user can select multiple options of interest. When the option is changed, only the opinions that match the criteria are presented in the visualization.	16
3.5	Distribution of opinions on the Retirement issue used for the study.	21
3.6	Distribution of opinions on the Game Addiction issue used for the study.	22
3.7	Distribution of participants' rating on the quality of the opinions used for the study. While participants were favorable to the opinions in general, participants felt opinions on Game Addiction issues were less agreeable and less novel.	30
4.1	Overall user flow of DataDive. (A) While reading, the user selects a statistical statement of interest to start exploring contextualization. (B) On the side, DataDive generates and presents a set of recommendations suggesting different potential contexts around the statistical statement. (C) Upon selecting a recommendation, DataDive matches the most relevant data from its database and visualizes the data. The user can use dropdown widgets to explore data in different contexts.	38
4.2	Three ways to start exploring contexts in DataDive. The user can (A) click on underlined statistical statements, (B) select the part of the text by themselves, or (C) type in their own statement of interest.	41
4.3	(A) UI for browsing the list of recommended questions and exploring data. (1) The first question is to verify the sentence with data. Each recommendation consists of (2) a question suggesting a different context and (3) the values of statistical indicators/entities/time periods. (B, C, D) After selecting one of the question from (A), DataDive provides (B) dropdown widgets to change selected statistical entities, indicators, and time periods, (C) visualization of selected data, and (D) metadata on data definition and source information.	42

4.4	The pipeline starts by parsing a sentence into a tuple (entity, indicator, date) and generating related recommendations. These recommendations are ranked and displayed according to the ranks. User selection of a recommendation triggers a match with relevant data fields in top-k datasets. Finally, the system filters and presents the pertinent data.	43
4.5	Example of parsing an input sentence into statistical entity, feature and indicator, and date.	44
4.6	Boxplot of the average rating of recommendations and the matching quality from the pipeline. (A, B) The results show that evaluators considered the recommendations meaningful, diverse, and interesting. (C) While evaluators were generally favorable to the matching quality, matching for the country had the most errors.	48
5.1	Conceptual diagram of FrameIt. FrameIt supports engaging in respectful online discussions by taking into account the moral foundations of the other discussion participants. From each comment, FrameIt identifies the moral foundations (e.g., Care/Harm) to help users better understand the perspectives of others. When writing comments, FrameIt provides feedback and suggestions that encourage users to consider the moral values of others.	60
5.2	FrameIt presenting the comments in the discussion thread. (A) FrameIt presents a list of moral foundations relevant to the comment as tags. Foundations with high confidence have a solid outline, while those with low confidence have a diffuse outline. (B) When hovering on one moral foundation, FrameIt explains how it is embedded in the comment as a tooltip. FrameIt also highlights sentences related to that foundation. (C) When the highlight button is clicked, FrameIt highlights each sentence with its moral foundation. When hovering on each sentence, FrameIt explains how each sentence is related to the moral foundation.	65
5.3	Reply screen of FrameIt. FrameIt supports users to express their opinions in online discussions while considering the moral foundations of the original poster to support respectful online discussions. (A) FrameIt shows the moral foundations of the comment the user replies to, both at the comment and sentence levels. (B) The user can select the target moral foundations based on the foundations of the original comment. (C) While writing the comment, FrameIt analyzes the comment in real-time to show whether the comment is aligned with the selected moral foundations. (D) FrameIt provides suggestions and feedback to write the argument.	66
5.4	User interaction of FrameIt for using suggestions on discussion points and feedback. (A) The user can write their comment in the textbox. (B) When the user clicks Suggest Discussion Points, FrameIt generates a list of potential discussion points based on the original post's content, the current comment content, and the target moral foundations. (C) When the user clicks the Generate Feedback button, FrameIt generates a list of feedback on logical flaws, misalignment of moral foundations, and the tone of the comment.	67
6.1	In the community profile page, DecideTogether shows the characteristics of the community as tags across four dimensions: Purpose, Participants, Stakes, and Tone. The user can further edit the community profile.	93

6.2	New Rule page of DecideTogether. (A) The user can elaborate on their initial moderation intent as the rule title. (B) The user can select the relevant community profile tags. (C) Based on the user’s moderation intent and community contexts, DecideTogether generates a draft of the rule text.	94
6.3	Rule editor page of DecideTogether. (A) DecideTogether shows the rule text on the left. (B) On the right, DecideTogether shows the automoderator logic constructed from the rule text. (C) Hovering over each logic item highlights the specific part of the text it is relevant to. Furthermore, it shows the number of incorrectly flagged and missed instances to help users assess the accuracy of the automoderator.	95
6.4	Moderation queue of DecideTogether. (A) When the user is overriding DecideTogether’s decision, the system elicits the reasoning. (B) For each automoderator logic, DecideTogether shows the reasoning behind the decision.	96
6.5	Interaction for editing the rule text on DecideTogether. (A) When the user edits the rule text, DecideTogether automatically shows (B) a preview on the logic update and (C) how the new logic would affect the previous automoderator decisions.	97
6.6	Interaction for getting suggestions to update the automoderator logic on DecideTogether. (A) DecideTogether shows a summary of each rule update. DecideTogether further shows (B) the effect of the fix on the rule text, (C) automoderator logic, and (D) the previous automoderator decisions.	98
6.7	Comparison of the quality of automoderator logic generated with and without contexts. Absolute scores from With-context and Rule-only conditions were comparable, but With-context conditions won in pairwise comparisons.	103
6.8	Per-subreddit difference in overall accuracy before and after adding the community context to the automoderator logic. While adding context improved accuracy for communities with underspecified rules, such as r/science and r/depression, it reduced accuracy for communities with stricter norms, such as r/AskHistorians and r/legaladvice.	105
6.9	Per-subreddit difference in overall accuracy before and after updating the automoderator logic from its errors. While updating the logic increased accuracy for most subreddits, it reduced accuracy for communities with already complex logic, such as r/AskHistorians.	105

Chapter 1. Introduction

1.1 Informed Opinions

Social issues increasingly require individuals to engage with complex, contested, and rapidly evolving bodies of information. Questions surrounding social issues, such as public health [1], climate change [2], online governance [3], artificial intelligence [4, 5], immigration [6], and social equity [7], rarely admit simple factual resolution. Individuals must interpret claims situated within broader social, political, moral, and epistemic contexts. Forming a considered position on such issues involves not only acquiring information but also examining how it should be understood, what assumptions underlie it, what values it reflects, how it relates to competing perspectives, what implications follow from accepting or rejecting it, and ultimately deciding on one’s own position.

I refer to such a considered position as an **informed opinion**: a position on a social issue that is grounded in relevant evidence, sensitive to alternative perspectives, and supported by reasons the holder can articulate. The importance of such positions has long been emphasized in deliberative democracy, which treats them as a normative prerequisite for legitimate collective decision-making [8, 9, 10]. While deliberation is ultimately a social process, it depends on **analytic work**—building an information base, weighing values, considering alternatives—that individuals carry out even within group settings [11]. Goodin [12] terms this “deliberation within” and argues that much of the reflective work attributed to group deliberation actually happens inside individuals as they privately imagine and weigh perspectives, a claim supported by empirical evidence that participants’ attitudes shift more during individual reflection than during group discussion [13]. This dissertation focuses on that individual analytic moment: the cognitive work people do when they encounter and reason about information on social issues. Such encounters take many forms across the process of forming an informed opinion — surveying the landscape of perspectives on an issue, examining a particular claim or argument, or articulating one’s own emerging position. In contemporary society, much of this encounter happens through online platforms, such as news sites, social media feeds, and discussion forums, where individuals engage with information about social issues privately and at scale.

1.2 Challenges in Forming Informed Opinions

A growing body of work in political psychology and cognitive science suggests three distinct obstacles to this analytic work, each pointing toward a different aspect of the same design opportunity.

First, individuals are often not motivated to engage in effortful reasoning. Humans conserve cognitive effort by relying on heuristics rather than systematic evaluation [14, 15], and even when effortful reasoning is engaged, it tends to be directed by prior commitments. People selectively evaluate evidence to reach conclusions they prefer [16, 17] and process information in ways that protect their standing within groups they identify with [18, 19]. Mercier and Sperber [20] argue that these patterns reflect the argumentative function of human reasoning rather than a defect to be eliminated, suggesting that biased individual reasoning is best addressed not by debiasing but by appropriate scaffolding.

Second, even individuals who are genuinely motivated to understand an issue deeply face a self-generation problem. Making sense of a social issue takes many forms: weighing competing perspectives,

examining a specific piece of information, or integrating it into one’s own point of view. Each of these requires considering perspectives that might disagree with the information at hand, the assumptions and values that underlie it, and the concrete situations in which it would apply. Some of this material lies outside the information at hand, in alternative arguments and considerations the reasoner has not encountered [21, 22]. Some lies within it, in framings and commitments that remain implicit [23, 24]. In either direction, people often fail to generate or identify such context on their own. Judgments are anchored on what is immediately available, with relevant counterfactuals and reference classes left ungenerated [25, 26], and reasoners construct coherent accounts from the information at hand without registering what is missing [27]. The limitation is not motivation but the cognitive reach of unaided recall and introspection.

Third, the current information environment does little to supply this context and often actively strips it away. Algorithmic curation favors fast, emotionally arousing content over reflective comparison [28, 29] and presents information stripped of the perspectives, assumptions, and comparative grounds that would render it interpretable. Critically, providing more information does not by itself produce more informed opinions: increased exposure to opposing views on social media reinforces rather than moderates prior positions [30], and higher numeracy amplifies rather than reduces polarization on contested empirical claims [2]. The bottleneck is not access to information but the interpretive resources available at the moment of reasoning.

Taken together, these three challenges point to a specific design opportunity. Because reasoners are unlikely to undertake demanding analytic work on their own, because the relevant context does not arise spontaneously even when they try, and because the surrounding information environment does not supply it, support must be provided at the moment of reasoning itself. Interactive systems are well-positioned to meet this need: interactive systems can intervene *at* that moment, offering interpretive support in context, on demand, and tailored to the specific information the reasoner is engaging with.

1.3 Interpretive Scaffolds as Facilitators of Individual Reasoning

This dissertation explores how interactive systems can support individual reasoning about social issues through what I call **interpretive scaffolds**: design elements that mediate the moment of encounter between a reasoner and a piece of information, surfacing alternative ways to interpret, examine, and integrate that information into the reasoner’s thinking. Interpretive scaffolds do not aim to replace the reasoner’s judgment or to deliver correct conclusions. Rather, they aim to make available the kinds of cognitive resources that individuals tend not to generate spontaneously during reasoning.

In this dissertation, I specifically focus on two types of interpretive scaffolds, corresponding to the direction of interpretation relative to the information under consideration.

Enriching scaffolds draw the reasoner’s attention outward, to elements of the broader information space beyond the immediate text. They may surface alternative arguments or perspectives the reasoner has not encountered, or additional factual context that may alter the evaluation of a contested claim. Enriching scaffolds can function as a *nudge*, prompting the reasoner to look beyond their current informational position, or as a *lens*, providing comparative grounds for critically assessing the focal information. Enriching scaffolds primarily address the second and third challenge above: when the surrounding environment fails to supply relevant context, the scaffold brings that context into view.

Restructuring scaffolds, by contrast, draw the reasoner’s attention inward, surfacing facets already latent within the given piece of information. They may function as a *map*, rendering visible the argumentative or value structures on which a claim rests, or as a *probe*, exposing assumptions that would otherwise remain implicit. By identifying and surfacing these latent facets, restructuring scaffolds help the reasoner grasp not only what the information claims but also on what basis. Restructuring scaffolds primarily address the second challenge: when the analytic work of unpacking a claim does not occur spontaneously, the scaffold externalizes the structure that would otherwise need to be constructed.

Both modes share a common stance toward the first challenge. Rather than attempting to debias individual reasoning, interpretive scaffolds work with the argumentative grain of human cognition [20]: they place the reasoner in the position of evaluating externally provided perspectives, structures, and comparisons, a stance at which humans are markedly more capable than at critically examining their own positions.

Critically, interpretive scaffolds are not merely items of information to be passively delivered. Their effectiveness depends on how they are elicited from underlying material, surfaced at the moment of information consumption, designed to invite engagement, and integrated with the reasoner’s ongoing thought.

1.4 Instantiations of Interpretive Scaffolds

I instantiate interpretive scaffolds through four interactive systems, organized along two dimensions: the two scaffold modes, enriching and restructuring, and three components of informed opinion formation that the systems support, engaging with diverse perspectives, consuming information critically, and articulating an informed position. *StarryThoughts* and *FrameIt* support engagement with perspectives through enriching and restructuring scaffolds, respectively; *DataDive* supports critical consumption through enriching scaffolds; and *DecideTogether* supports articulation by combining both modes.

StarryThoughts supports the exploration and respectful comprehension of diverse opinions by surfacing the demographic and ideological landscape of perspectives on a social issue as **enriching scaffolds**. Where readers ordinarily encounter opinions in environments that limit exposure to differing perspectives, *StarryThoughts* visually externalizes the broader space of opinion diversity, situating each opinion within the demographic and ideological diversity of those who hold it. This allows readers to locate their own stance within that space and to engage with viewpoints they would not have sought out unaided.

DataDive supports critical assessment of a statistical claim by surfacing relevant external data, along with guidance on what to compare and how, as **enriching scaffolds**. Statistical statements in everyday discourse tend to be evaluated in isolation from the broader empirical context that would qualify or recontextualize them. *DataDive* surfaces relevant external datasets and supports the reader in comparing them against the focal statistical claim, broadening the range of evidence the reader encounters and enabling more in-depth evaluation of the information at hand.

FrameIt supports respectful engagement across disagreement by surfacing the moral foundations embedded in online opinions as **restructuring scaffolds**. Where readers might otherwise engage only with the surface argument of an opinion, *FrameIt* makes the underlying moral foundations visible, rendering the argument’s value basis available for inspection. This allows readers to evaluate not only what an opinion asserts but the deeper commitments that make it compelling to those who hold it, and to share their own views while respecting those of others.

DecideTogether extends the construct to a practitioner setting, online community moderation. Moderators should translate their community-specific moderation intents into concrete rules, which requires situating those intents within the community’s lived context and examining the latent assumptions underlying them. DecideTogether brings both modes of scaffolding together to support this work: it provides analyses of community context as well as actual contributions from the community as **enriching scaffolds**, and system-generated interpretations of moderation intent, expressed in human-readable rule text and machine-executable automoderator logic, as **restructuring scaffolds**. By presenting the system-generated interpretation of moderation intents and how a moderator’s intent applies to real posts and communities in action, the system supports iterative reflection on and refinement of that intent into a concrete rule text.

Across these four systems, I examine not only whether interpretive scaffolds support better understanding and deeper reasoning on the issue, but also how they must be designed: how they are elicited from underlying material and surfaced at the moment of reasoning to be integrated into the reasoner’s ongoing thoughts.

1.5 Contributions

This dissertation makes three contributions to the study of how interactive systems can support informed opinion formation on contested social issues.

First, I introduce interpretive scaffolds as a design construct for supporting individual reasoning at the moment of information encounter. I motivate the construct from three challenges to informed opinion formation, and I distinguish two modes—enriching and restructuring—that target different challenges. The construct is offered as a lens for organizing the design of the systems in this dissertation and for informing future work on systems for reasoning support.

Second, I contribute design strategies and computational techniques for realizing interpretive scaffolds in practice. Through four systems—StarryThoughts, DataDive, FrameIt, and DecideTogether—I show how scaffolds can be elicited from external information sources and from the information of focus itself using large language models and structured extraction, surfaced at the moment of reading without disrupting it, and integrated with the reasoner’s ongoing thought.

Third, I provide empirical evidence from user studies of each system, characterizing how interpretive scaffolds shape reasoning across four tasks: engaging with diverse opinions, assessing statistical claims, examining the value basis of arguments, and translating abstract moderation intent into rules. I synthesize these findings into an account of how interactive systems can support individual analytic work on contested social issues.

Thesis statement: Interpretive scaffolds foster understanding of complex social issues by surfacing latent facets from unstructured information and enriching it with broader contexts, contributing to respectful engagement with diverse opinions, critical consumption of information, and articulation of one’s own informed position.

1.6 Thesis Overview

- Chapter 2 reviews the theoretical background and prior work across four areas foundational to this thesis: (1) Normative theories around informed opinion formation, (2) Cognitive challenges in forming opinions, and (3) Interactive systems for supporting information engagement.

- Chapter 3 presents **StarryThoughts**, an interactive system that supports the exploration of diverse perspectives on social issues with a respectful attitude towards different opinions, leveraging the demographic profiles as the external interpretive scaffold.
- Chapter 4 presents **DataDive**, an interactive system to support a deeper understanding of statistical statements with questions and an external dataset.
- Chapter 5 presents **FrameIt**, an interactive system that supports better comprehension of different opinions by exposing the moral foundations embedded in them.
- Chapter 6 presents **DecideTogether**, an interactive system supporting articulation of a solution for a social problem
- Chapter 7 summarizes the lessons from designing and evaluating interactive systems for interpretive scaffolding to discuss main design considerations, potential tradeoffs, and the limitations of this approach to guide future researchers.

Chapter 2. Background

2.1 Informed Opinion and its Formation

In this section, I review the literature on informed opinion and its formation, its role in democratic decision-making, and the normative and cognitive accounts of how individuals arrive at such opinions.

2.1.1 Informed Opinion in Democratic Theory

Informed opinion is a normative prerequisite for democratic decision-making. By “informed opinion,” I refer to a citizen’s position on a social issue that is grounded in relevant evidence, sensitive to alternative perspectives, and supported by reasons the holder can articulate. Forming such an opinion requires exploring relevant information, considering alternative perspectives, and weighing them against the values at stake to construct a defensible position.

The importance of informed opinion is most explicit in deliberative democracy. Deliberative theorists reject the aggregative view that democratic legitimacy derives from the mere counting of preferences [9, 31, 10]. On the aggregative view, preferences are taken as given; on the deliberative view, they must be formed and revised through reasoned exchange. Legitimacy depends not only on what citizens decide but on how they arrive at their decisions [32, 33, 34]. Uninformed or unreflective preferences, even when aggregated through fair procedures, cannot confer legitimacy in the same way. Habermas [8] grounds this in communicative action: legitimate opinions are those that could be defended with reasons to others under conditions free from coercion. What makes an opinion “informed,” on this view, is not only its factual basis but its responsiveness to reasons others could accept [9].

2.1.2 The Individual Work of Deliberation

Deliberation is often described as a fundamentally social process, but this framing obscures the individual cognitive work it requires. Black et al. [11] distinguish two processes in deliberation: an analytic process—building an information base, prioritizing values, identifying and weighing solutions, and reaching decisions—and a social process—comprising norms of balanced participation, mutual comprehension, consideration, and respect. While Black frames both as discussion behaviors, a parallel line of work has argued that much of the analytic work deliberation takes place within individuals rather than between them. Goodin [12, 35] terms this “deliberation within,” arguing that the internal-reflective aspect of deliberation must ultimately take place within each individual, even in collective settings. Empirical work supports this: Goodin and Niemeyer [13] found that jurors’ attitudes changed more during the information-gathering phase, which involved substantial individual reflection, than during the formal group discussion. Pingree [36] theorizes the mechanism, showing that the anticipation of expression and the act of composing one’s position produce cognitive effects on the reasoner independent of social exchange.

2.1.3 How Individuals Form Informed Opinions

Two broad accounts of how individuals form informed opinions coexist in the literature. The normative account, drawn from deliberative theory and classical epistemology, treats opinion formation as a deliberate process of seeking information, evaluating evidence, considering counterarguments, and updating beliefs accordingly [8, 9, 37]. On this view, the individual approximates an ideal reasoner: open to revision, attentive to opposing views, and motivated by the better argument.

The descriptive account, drawn from political psychology and cognitive science, complicates this picture. Citizens routinely rely on heuristics and informational shortcuts rather than systematic evaluation [38], and what appears as a stable “opinion” often reflects whichever considerations are momentarily accessible rather than a settled position [39]. The next section examines these departures in detail and argues that they point to a specific design opportunity.

2.2 Challenges with Forming Informed Opinion Formation

Forming informed opinions about social issues is a demanding interpretive task. Such issues are characteristically ill-structured: they involve contested values, incomplete and conflicting evidence, and disagreement over which framings and trade-offs are relevant [40, 41]. Reasoning about them thus requires not the retrieval of a correct answer but the active interpretation and weighing of heterogeneous considerations. This section reviews three challenges that make this interpretive work difficult in practice: motivation, cognition, and environment. Also, this section argues that the gap is not primarily about access to information but about the interpretive resources available at the moment of reasoning.

First, individuals are often not motivated to engage in the effortful reasoning that informed opinion formation requires. Humans tend to conserve cognitive effort by relying on heuristics rather than deliberate reasoning, a contrast captured by dual-process accounts of cognition [14, 15]. Even when effortful reasoning is engaged, it is often directed by prior commitments rather than open inquiry: people selectively evaluate evidence to reach conclusions they prefer [16, 17], and process information in ways that protect their standing within groups they identify with [18, 19]. Mercier and Sperber [20] argued that these patterns reflect the function of human reasoning, which evolved for argumentation rather than solitary truth-seeking, suggesting that biased individual reasoning is not a defect to be eliminated but a tendency to be channeled through appropriate scaffolding.

Second, even when individuals are motivated to reason carefully and have access to relevant information, the interpretive context necessary for evaluating claims rarely becomes cognitively available on its own. Here, interpretive context refers to the surrounding perspectives, assumptions, comparisons, and application conditions through which a claim becomes meaningful and evaluable. Some of this context lies outside the information at hand, in alternative arguments and comparative cases the reasoner has not encountered, and some lies within it, in framings and commitments that remain implicit [23, 24]. In either direction, people rarely generate or identify such context on their own. Even when explicitly prompted, individuals struggle to produce alternative theories and counterarguments to their own positions [21, 22], and they often fail to spontaneously coordinate across them to construct an integrated interpretation [42]. Judgments are anchored on what is immediately available, with relevant counterfactuals and reference classes left ungenerated [25, 26], and reasoners construct coherent accounts from the information at hand without registering what is missing [27]. This limitation extends from abstract text to situated application, where the latent implications of a principle remain invisible until brought into

contact with concrete cases [43], and where probing for mechanism often reveals that confident positions rest on shallower understanding than the reasoner had assumed [44, 45]. The broader evaluative context is typically absent at the moment of encounter, not because individuals are unwilling, but because human cognition struggles to actively construct it without external prompting.

Third, the current information environment does little to supply this context, and often actively strips it away. Early commentary emphasized the filter bubble effect, in which algorithmic curation narrows the range of viewpoints users encounter [46, 47]. Subsequent empirical work has shown that algorithmic filter bubbles are more modest than initially claimed, with users on major platforms encountering substantial cross-cutting content, though still suffering from selective exposure within homophilous networks [48, 49, 50]. More fundamentally, algorithmic curation shapes the interpretive context in which information is encountered: it favors fast, emotionally arousing content over reflective comparison [28, 29] and presents information stripped of the perspectives, assumptions, and comparative grounds that would render it interpretable. Providing more information, by itself, does not produce more informed opinions: increased exposure to opposing views on social media has been shown to reinforce rather than moderate prior positions [30], and higher numeracy can amplify rather than reduce polarization on contested empirical claims [2]. The bottleneck is not access to information but the interpretive resources available when that information is consumed.

Taken together, these three challenges point to a specific design opportunity. Because individuals are not naturally inclined to engage in open-ended reflective reasoning, because the interpretive context necessary for evaluating claims rarely becomes cognitively available on its own, and because the surrounding information environment often fails to provide such context, support for informed opinion formation must be embedded directly into the moment of information consumption itself. The systems presented in this dissertation pursue this through *interpretive scaffolds*: design elements that surface the perspectives, assumptions, comparisons, and application contexts relevant to a piece of information as the reasoner engages with it. By externalizing interpretive operations that people rarely perform spontaneously, lowering the activation barrier for reflective reasoning, and making otherwise latent comparisons cognitively available at the moment of judgment, such scaffolds aim to support more reflective and contextually grounded opinion formation.

2.3 Interactive Systems for Supporting Information Engagement

A substantial body of HCI research has explored how interactive systems can support people in engaging more deeply with information. Although these systems were rarely framed as interpretive scaffolds, many of them can be seen as design interventions that mediate the moment of encounter between a reader and a piece of information. In this section, I organize prior systems along the two modes introduced in Chapter 1—*enriching* and *restructuring*—and identify the gaps that motivate the four systems presented in this dissertation.

2.3.1 Surfacing the broader information space

Reasoning about social issues depends on comparison with a broader information space—alternative perspectives, related evidence, and the voices of people affected by the issue. In ordinary information environments, this broader space is not automatically visible to the reader. One line of HCI research

supports interpretation by bringing parts of this external space into view. From the perspective of enriching scaffolds, this work falls into three groups.

The first group externalizes opinions and perspectives that the reader has not encountered. Earlier approaches [51, 52] focus on surfacing the existence of diversity of perspectives. ConsiderIt [53], Balancer [54], and Burst Your Bubble [55] visualize users’ information consumption biases and nudge toward balanced information consumption. More recent work re-ranks or restructures discussion content to expose readers to a wider range of views [56, 57, 58, 59, 60, 61]. These systems act as a nudge: they push the reader to look beyond their current informational position.

The second group brings in external evidence and information to contextualize a claim. Factful links statements about public budgets to external data [62]. Contextifier and NewsViews visualize the time-series or geographic data behind news stories [63, 64], and VizByWiki and Silva et al. extend this to general news articles [65, 66]. Work on academic reading, such as ScholarPhi [67], Paper Plain [68], CiteRead [69], and Threddy [70], uses a similar mechanism, pulling in external information to support comprehension of difficult text. A related body of work re-expresses unfamiliar numbers in terms more familiar to the reader, so that the magnitude of a claim can be assessed against known reference points [71, 72, 73, 74]. These systems act as a lens: they provide the reader with a point of comparison against which to evaluate the claim at hand.

The third group treats not the claim itself but the people behind it as the external information. Saveski et al. expose users to feeds from political opponents to humanize them [75], and Nelimarkka et al. study how revealing partisan affiliation shapes engagement [76]. Other systems surface lived experience behind a position through personal narratives [77] or stakeholder profiles [78]. In discussion threads, source-position indicators [79, 80] and tension indicators [81] make the origin and conflict structure of opinions visible. These systems provide a lens by attaching identity and context to opinions, so that readers can interpret what is said in light of who is saying it and why.

2.3.2 Making Latent Structure Visible

A second line of work focuses on making the structure already present in a text visible. From the perspective of restructuring scaffolds, this work uses two mechanisms: showing the latent structure itself or prompting the reader to examine it through probing questions.

The first mechanism externalizes the structure of a text in an alternative representation. Charagraph and Elastic Documents generate charts from data-rich text so that the quantitative structure of a passage becomes visible alongside the prose [82, 83]. Papeos links a paper to the author’s presentation video, surfacing the argumentative structure across modalities [84]. Work on coupling text and visualization within documents follows the same idea at finer granularity, linking phrases to chart elements so that readers can move between textual and visual representations of the same content [85, 86, 87]. A separate tradition of argument visualization renders the claim-evidence-warrant structure of arguments explicitly, so that the reasoning behind a position can be examined rather than only read [88, 89, 90, 91]. Baumer et al. [92] explored the impact of visualizing political framing from texts. These systems treat the text as something whose internal structure can be drawn out and inspected.

The second mechanism prompts the reader to examine the latent structure themselves through guiding questions. CritRainer and CReBot generate questions to support critical reading of academic papers [93, 94], while ReviewAid and Marvista take a similar approach for news articles [95, 96]. AngleKindling reframes press releases into alternative angles a reporter might pursue [97]. Tanprasert et al.’s debate chatbot prompts users to re-examine positions formed from watching videos [98]. Lim et al.’s

browser extension highlights fallacious passages and offers questions for further inquiry [99]. Wang et al. take a related approach in the context of moral reasoning, building a discussion platform where users annotate news articles with moral labels and self-report the moral foundations behind their positions, thereby supporting mutual understanding across viewpoints [100]. Rather than drawing the structure out for the reader, these systems provide the reader with probes to find it.

2.3.3 Combining Enrichment and Restructuring

Some interpretive work cannot be done by examining a single text. Forming a policy, deciding how a community should be governed, or making sense of an unfamiliar domain begins from an intent that is underspecified. The reasoner knows roughly what they want, but its concrete shape emerges only through contact with specific cases. Recent HCI work supports this articulation by combining enrichment and restructuring.

One thread supports individual sensemaking. Sensecape [101] and Graphologue [102] let users externalize the structure of their understanding on a canvas, while LLMs generate content that extends it. ChoiceMates [103] and Selenite [104] elicit the dimensions along which options can be compared and generate alternatives along them. The user’s understanding is not stated in advance and applied; it is shaped through interaction with generated cases and structures. A parallel thread supports policy articulation. PolicyCraft [105] helps community members develop moderation policies by surfacing cases the policy would apply to, allowing refinement in response. Botender [106] supports the construction of moderation bots through iteration on rules against example posts. A parallel line of work on LLM alignment shares the same mechanism. ConstitutionMaker [107] converts a user’s case-by-case feedback on chatbot outputs into natural-language principles that steer subsequent behavior. Policy Projector [108] lets users define LLM behavior policies by mapping concepts and cases onto the same plane, refining policies against the cases they would apply to. EvalLM [109] and EvalGen [110] support the iterative refinement of evaluation criteria for LLM outputs, with EvalGen explicitly identifying criteria drift: users cannot fully specify their criteria in advance, and refine them as they encounter outputs. Across these systems, abstract intent — whether a moderation policy, a chatbot principle, or an evaluation criterion — is treated not as something to be specified up front, but as something to be drawn out through interaction with concrete cases.

Both threads share a mechanism. External cases or alternatives are introduced (enrichment), and the reasoner’s intent is re-expressed against them — surfacing assumptions, exposing gaps, and prompting refinement (restructuring). Through the iterative process, the user reflects on their intent and better shapes it.

Chapter 3. StarryThoughts: Facilitating Diverse Opinion Exploration on Social Issues

This chapter presents StarryThoughts. This chapter has adapted, updated, and rewritten content from a paper published at ACM CSCW 2021 [111]. All uses of “we”, “our”, and “us” in this chapter refer to the coauthors of the aforementioned paper.

3.1 Motivation & Contributions

The deliberative ideal of democracy suggests that the main virtue of democracy is the free and equal exchange of arguments in a respectful manner [112, 37]. Online spaces have shown the possibility of achieving this ideal by facilitating information and opinion exchanges between people [113, 114, 115]. However, at the same time, the current designs of online spaces are responsible for aggravating the problem of political polarization by providing places where like-minded people can come together to form networks. As these networks come together, individuals begin to form group identities. Ingroup-outgroup relationships are formed and, as people form stereotypes and biases towards the outgroups, interaction between groups becomes limited and results in a further polarized online space [116]. Online spaces also provide methods to filter content depending on users’ preferences, which is known to be a cause of the filter bubble [46], where personalized and algorithmic filtering limits access to diverse information, thereby creating an invisible barrier that presents a biased worldview for the user. In this way, polarization also hinders democratic communication by prohibiting each person from being exposed to various alternative ideas [34, 117]. These barriers hinder communication and, in turn, prevent the achievement of an ideal deliberative and democratic space.

Previous work on mitigating polarization has mainly focused on increasing people’s awareness of diverse opinions on an issue. Common approaches include explicitly presenting the bias of the user [54, 55, 53] or showing the diversity of opinions [56, 51]. However, simply knowing the existence of diverse opinions might not be enough. Prior research suggests that the mere exposure of diverse opinions might not immediately mitigate polarization [30]. Deliberation literature suggests that listening to other opinions with respect is essential for understanding different points of view [37, 118], but designs for fostering respect towards diverse opinions have been relatively less explored. One of the few examples is Reflect [119], which introduced interaction designs based on active paraphrasing to promote listening in online discussions.

In this work, we aim to explore the possibility of using demographic identities to foster respect towards people with different opinions in the online space. We believe that the identities of individuals could provide context for each opinion to support the user’s role-taking process [118] and judgment of opinion credibility. We also use the identity information as a method to bring attention to stereotypes that a user might have, as well as to increase their perceived variability of outgroups, ultimately mitigating outgroup stereotypes and facilitating democratic conversation [120].

We propose StarryThoughts, an online system that supports users to explore diverse perspectives on social issues. The system primarily aims to present the diversity of opinions as well as the context behind the opinions by introducing a visualization of opinions on a 2-D space, following the metaphor of “starry nights” (Figure 3.3). On top of that, the system provides several different support mechanisms

to help the users explore the opinion space. First, the system enables users to perceive the diversity of backgrounds and experiences behind the opinions by presenting the demographic information of the author of each opinion. The system supports the users in understanding the perspectives of a specific social group by filtering the opinions by the demographics of the authors. Second, using the demographic information of the authors, the system explicitly enables the user to state and test their expectations on the demographic identities of the supporters and the dissenters of the issue. Finally, the system engages users to explore the broader range of opinions in the 2-D space by visually recommending opinions with the least semantic similarities to guide them to engage with opinions that contain different perspectives.

We conducted a lab study to investigate how participants would use the features of the system to explore a large number of opinions and how exploring diverse opinions might affect participants' opinions on social issues and their attitudes toward disagreeing opinions. Study results with a total of 56 participants showed that with StarryThoughts participants explored a diverse range of opinions with a 2-D visualization of opinions and presentation of demographic identities of the opinion authors. The difference in participants' opinions, level of confidence in their opinions, and argument repertoire before and after using the system showed that StarryThoughts helped the participants to express more concrete opinions and recall more arguments around the issues.

The contributions of this chapter are as follows:

- Design implications for systems facilitating the diverse exploration of opinions and fostering a respectful attitude toward different opinions;
- StarryThoughts, an online platform for sharing and exploring a diverse range of opinions on social issues;
- Findings from a user study showing that with StarryThoughts participants explored a diverse range of opinions with respectful attitudes towards them.

3.2 Initial Design & Iteration

We designed an initial version of the system with the primary goal of exposing the diversity of opinions. (Figure 3.1). The system aimed to provide a visual platform that users could freely navigate among various opinions, stances, and identities. This version was designed to help users understand the variability and diversity of opinions (through the navigation of space), as well as to emphasize the multifaceted relationship between identity factors and opinions. Table 3.6 shows the design elements of the system compared to existing approaches. Below, we elaborate on the specific design elements we used to implement these goals on an interactive system.

When the user enters the system, the user is presented with a 2-D spatial visualization of opinions that resembles a “starry night sky of opinions”. We used a 2-D spatial visualization to present a visual impression of opinion diversity and the relationship of opinions and demographic identities. Compared to commonly used lists of opinions, the spatial visualization also removed the hierarchy of elements in organizing the opinions so that each element was as equally accessible as any other. Each opinion is visualized as a dot on the screen. Unlike previous research that placed the opinions by the similarity of opinion profile [51], the horizontal position of each dot was determined by the author's stance on the issue. By doing so, we expected the users to see the opinion distribution easily. While each opinion has information on the author's identity, we intentionally did not encode any information in the vertical axis

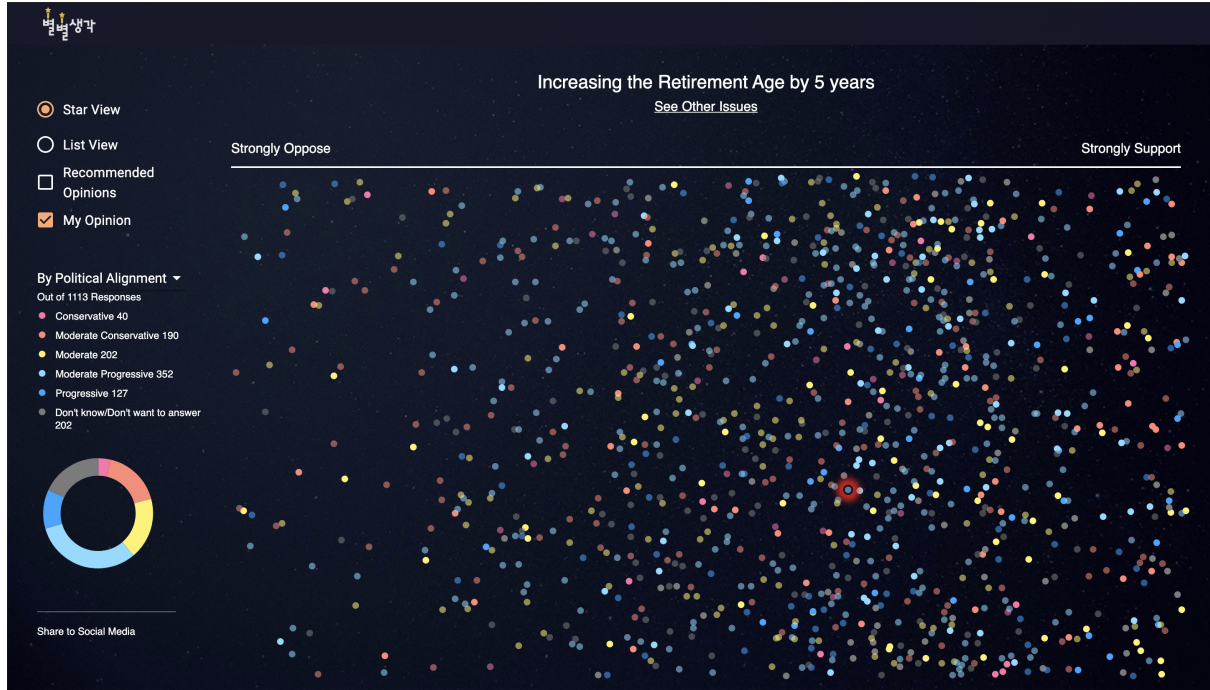


Figure 3.1: The initial version of the system. The initial version of the system presents the variability and diversity of opinions by plotting them as dots in a 2-D space.

to not give an impression that each social group is segregated. Instead, we placed each opinion randomly on the vertical axis so that each opinion could be equally noticed. The data points are color-coded with regard to which demographic group the author is in. Users can choose between three modes of color-coding: age group, gender, and political alignment, to see how the distribution of each demographic characteristic matched up within the opinion distribution.

One of the significant design considerations was deciding which and how many identity factors to disclose. If the factors were too few or too general, they might not have enough informational value. Fine-grained identity factors would provide the readers with a more detailed context for the opinions. However, it can also lead to various problems, such as an increased level of stereotyping of or prejudice against the authors, or the possibility of disclosing personal information. With such considerations, we chose a combination of four general categories for the users to classify themselves in: age group, gender, occupation, and political stance. These were criteria that were already being used widely for public opinion polls, so they are easy to understand and general enough not to pose a threat to the users' privacy while still providing meaningful context.

The user can filter the opinions by selecting a demographic to focus on, so that they are able to choose and focus on the opinion distribution of a particular demographic group. By doing so, the user could investigate the opinion distribution of specific groups. For example, the user can check whether their peers share the same view on an issue or verify the validity of their preconceptions on specific groups. The user can hover or click on the dots to see the arguments of others. The user could also use the recommendation feature of the system to discover new perspectives. In the initial version, the user could see the opinions of the 'most distant' authors, who are the people with different stances and different demographic characteristics from the user.

We deployed the initial version of the system in collaboration with a nationwide broadcasting com-

pany in South Korea. The system was open from October 9th, 2019, to October 31st, 2019. During the deployment, 1,209 people visited the website, and 335 people signed up. A total of 1950 opinions and 144 open-ended arguments were shared on the platform, spanning over six controversial social issues.

From the experience of deployment and several pilots for design iterations, we came up with several design lessons to better help users explore diverse opinion landscapes. In the following subsections, we introduce some of the lessons (denoted as L1-L3) that led us to the second version of the system.

3.2.1 L1: Inspecting fine-grained demographic groups

In the initial version, the system supported filtering opinions with only one demographic characteristic at a time. For example, a user would select the ‘Male’ filter from the ‘Gender’ category to see all the opinions that male-identifying authors provided. However, pilot study participants mentioned that using only one demographic characteristic was too coarse-grained. We revised the system to support multiple filtering options simultaneously after receiving feedback on the lack of specificity in the group selection during pilot studies. Users can select multiple filters across the three categories to control the level of granularity and also the amount of information provided at a given moment.

3.2.2 L2: Leveraging users’ prior knowledge for engagement

From the pilot studies, we learned that one of the most common usage scenarios of the system was to discover the trend of opinions according to demographic factors. Specifically, the participants had prior expectations of the strongest supporters or dissenters of an issue in terms of demographics. They used the system to check whether such expectations were aligned with the opinion distribution in the system. The participants mentioned that discovering an unexpected trend of opinion to a specific demographic group or discovering opinions that are out of trends made exploration more intriguing and made them explore the opinions to understand the reasoning for such instances. Inspired by Kim et al. [121], we aim to support the users of the system to externalize their prior expectations and verify them.

3.2.3 L3: Active guidance for exploring diverse opinions

The users of the initial version of the system were only able to see the dispersed opinions, with support from a simple recommendation feature focusing on presenting opinions from different social groups. So, the initial algorithm suggested opinions that are the most distant from the user regarding the authors’ demographic profiles and their stance on the issue. However, from a series of pilot studies, we learned that the users heavily relied on the recommendation feature when they started exploring the opinions, as they tended to be overwhelmed by seeing a large number of data points all at once. The participants were also concerned that randomly selecting opinions would often return opinions too similar to what they had read before. Therefore, we decided to improve the recommendation by focusing on fostering exploration of more diverse arguments with more transparent criteria.

3.3 System

In this section, we introduce StarryThoughts, an online platform to support the exploration of a diverse range of opinions. Upon entering the system and adding their own opinion and demographic information, the user can see a 2-D visualization of opinions, arranged by the stance on the issue. Then, the user can freely explore the space and read individual opinions using various system features.

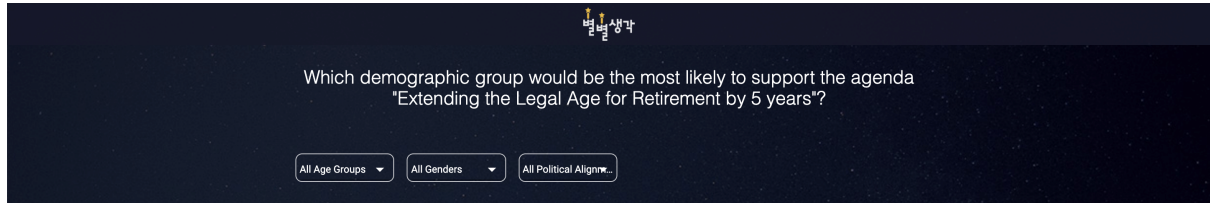


Figure 3.2: Prediction screen for the user. Before seeing the visualization of the opinions, the user is first asked to share their prediction on the groups who would likely support or object to the issue. The user can quickly verify their prediction by seeing actual opinions in the next step.



Figure 3.3: An overview of StarryThoughts. Each opinion is displayed as a dot in a two-dimensional space. The user's opinion is highlighted in red, and recommended opinions are highlighted in yellow, blue, or green according to the recommendation type. (A) With the filters, the user can specify demographic groups of interest. (B) The user can quickly test their predictions on who would support or oppose the issue with the quick filters. (C) By clicking or hovering on one of the dots, the user can read the opinion along with the demographic profile of the author.



Figure 3.4: Demographic filter of StarryThoughts. Each available option is presented as a dropdown menu at the upper left corner of the page. The user can select multiple options of interest. When the option is changed, only the opinions that match the criteria are presented in the visualization.

3.3.1 System usage scenario

To present how the user would explore diverse opinions on social issues with StarryThoughts, we introduce a representative scenario of using the system, constructed from the activity logs of the users.

Mina is a college senior, recently having been on the search for future jobs after graduation. She recently heard about a suggestion for a new policy that aims to increase the legal retirement age. In her opinion, the new policy would worsen youth unemployment as the firms would be less likely to hire people. She also thinks that a turnover in the workforce is necessary to promote innovative thinking. Still, she wonders what others would think about the policy and visits StarryThoughts.

Predicting and verifying supporters and dissenters of the issue (L2)

Before showing the opinion distribution, the system asks her to predict the demographic groups likely to agree or oppose the legislation (Figure 3.2). Mina immediately selects the age group of the 20s as likely to oppose the legislation since she supposes that most people in her age group would be in a similar situation to her own. She considers a little bit about groups that would support the legislation and selects ‘60s Male’, as she thinks they would be the ones whose jobs would be terminated if this came through.

She sees the ‘Star View’, where opinions are visualized as dots scattered on a 2-dimensional space (Figure 3.3). She notices many data points, but she is not sure which dot to choose first. She notices the ‘Likely Supporters’ and ‘Likely Dissenters’ filters and decides to take a look at them first (Figure 3.3B).

She clicks on the ‘Likely Supporters’ filter to see the responses from male people in their 60s. As she sees how the dots are dispersed, she realizes that the opinions are skewed to support the issue as she expected. At the same time, she notices that there are still some dots on the opposing side. She reads one of them, which says that increasing the retirement age causes more social costs than benefits and that the rights of the older citizens would be secured better when provided through social welfare policies. She is impressed by the opinion as she did not expect that he would support an idea against his own benefit. She leaves a Like on the opinion.

Exploring the trend of opinions per demographic group (L1)

After reading the opinion related to the issue with social welfare policy, she realizes that the solution of strengthening social welfare policies is not the one she had considered before. She also wonders whether people’s attitudes towards welfare policy would affect their opinion on the issue. She uses the demographic filters (Figure 3.3A) to check how the opinion distribution differs by political alignment (Figure 3.4). She observes that the opinions do not significantly differ by political alignment and concludes that the social welfare perspective was not a decisive factor.

Exploring diverse opinions with recommendation (L3)

She still wants to learn more about diverse arguments around the issue, so she goes back to see all the dots. Among the distribution of dots, she discovers that some dots in the system were glowing in different colors to indicate recommended opinions (Figure 3.3).

She clicks her opinion first to get a recommendation based on it, then considers the meaning behind the colors. The yellow dots are the opinions with “Opposite Stance, Different Point of View”. They support the policy because 60 is too low of a bar for people to give up their skills and abilities to work, which is an entirely different perspective from her own. The blue dots are the opinions with “Opposite

Stance, Similar Point of View”. Among them, she finds opinions supporting the policy because it would lead to a better workforce. From the green dots, described as “Same Stance, Different Point of View”, she discovers different arguments for opposing the policy, such as that the quality of life for older citizens would be better if they could retire and receive pensions earlier.

Mina wants to know more about the points related to welfare, so she goes back to the opinion that mentioned welfare and starts reading the opinions with blue highlights. One of them emphasizes the importance of enforcing social welfare policies to ensure citizens’ rights, but claims that keeping citizens in employment for longer is a good way to reduce needless budget losses in the social welfare sector. Mina is intrigued by how similar lines of reasoning are used in different contexts to create completely contrasting opinions. She is also surprised at the fact that this opinion is written by a man in his 20s, but after seeing that he is a graduate student, she justifies the idea that he might think that way because he would probably not be as involved in the job market (Figure 3.3C).

She reads a few more recommended opinions to find any new perspective on the issue. After exploration, she leaves the system, with her thoughts on the issue moved towards neutral.

3.3.2 Recommendation algorithm (L3)

From the design lessons, we decided to support users exploring the space by actively providing recommendations from new perspectives. Previous work on fostering exploration of diverse perspectives focused on recommending agreeable opinions. To do so, the metric for recommendation measured the diversity of the population the opinion appealed to [51, 53]. Some other research tried recommending attitude-inconsistent opinions using a classifier to predict people’s stance on the issue [55]. We aimed to encourage users to consider opinions from new perspectives. We assumed that opinions with similar perspectives would be semantically similar. Therefore, we opted to recommend opinions that were semantically dissimilar to the ones currently being read.

The system determines the distance between points of view using the document similarities computed with the method of Arora et al. [122]. In this method, document embedding is constructed by averaging the embedding vectors of the words in the document. While averaging, each word vector is weighted by the inverse frequency of the word in the corpus.

As the system was built and deployed in Korean for a Korean audience, the users’ opinions were also in Korean. Therefore, it was impossible to apply the technique directly due to the lack of reliable Korean word vectors. However, as the algorithm weights the words by inverse frequency, we expected that the algorithm would work reasonably if we could get the keywords in each opinion correctly. So, we decided to use the pre-computed English word vectors despite sacrificing some meaning of the text. To do so, we translated the collected opinions into English using Naver’s Papago API ¹ after correcting for spelling and grammar for better machine translation. After translation, we computed each opinion’s embedded vector with the algorithm using pre-trained GloVe word embeddings built upon 42B tokens from Common Crawl [123].

3.3.3 Implementation details

The front-end of StarryThoughts is implemented with React. The back-end is built with Node.js and served on AWS Lambda with a MongoDB database. For the recommendation algorithm, a separate

¹<https://papago.naver.com/>

Python server is used to compute the document similarities of the opinions and make recommendations based on them.

3.4 Evaluation

To understand whether the features of StarryThoughts effectively support users’ exploration of diverse opinions and how users’ opinions and attitudes towards the issue would be affected by using StarryThoughts, we conducted an exploratory user study. In evaluating ConsiderIt [53], researchers investigated how users used the system to explore diverse opinions and how such exploration helped their opinion formation. Opinion Space [51] attempted a controlled within-subjects user study to evaluate their system against baseline interfaces in terms of user engagement, helpfulness for finding useful opinions, diversity of opinions read by users, and agreement and respect towards different opinions. Inspired by the previous work, we developed the following research questions to thoroughly understand the experience and the effects of using StarryThoughts.

RQ1 How would the users of StarryThoughts use demographic identities of the authors and recommendations to explore diverse opinions on social issues?

RQ2 How would the experience of using StarryThoughts to explore opinions on social issues affect users’ opinion strength and quality?

RQ3 How would the experience of using StarryThoughts affect how the user evaluates diverse opinions around social issues?

3.4.1 Participants

Our user study consisted of two separate rounds. In the first round, we focused on understanding how using StarryThoughts would affect users’ opinions on social issues and attitudes to others’ opinions with pre-/post-surveys. The second round of user study was designed to thoroughly understand a) how users used each system component to explore opinions and b) how using demographic identities for exploring opinions affected users’ reaction to the opinions of others.

We had 50 participants for the first round. 27 of them were female, and 23 of them were male. Their ages ranged from 18 to 42, and the average age was 24.1. We recruited the participants by posting calls for participation in online communities of two universities in South Korea. The participants were compensated with KRW 10,000 ($\approx$ USD 8) for participating in the study. In the results, we refer to these participants as P1 through P50. Due to a glitch in the system, the system usage logs of 9 participants were lost, so we excluded them while analyzing the logs.

There were six participants for the second round, and all of them were male. Their ages ranged from 20 to 32, with an average age of 26.5. We recruited the participants via posting calls for participation on an online community of a university in South Korea. The participants were compensated with KRW 20,000 ($\approx$ USD 17) for their participation. In our results, we refer to these participants as Q1 through Q6.

3.4.2 Conditions

In this evaluation, we focused on understanding the experience of using StarryThoughts and whether the features of the system support exploring diverse opinions rather than quantitatively evaluating the

effect of the system against existing approaches. Therefore, we ran the user studies without a specific control condition.

We used two controversial socio-political issues widely known to the public in South Korea at the time of the study (May 2020). We chose the issues for which it is expected that people have different attitudes depending on their demographic identities. At the same time, we decided to refrain from choosing too controversial and sensitive issues so the users of the system would feel comfortable accepting and respecting the opinions of others.

The first issue is increasing the legal age for retirement (Retirement Age) [124, 125]. Due to population aging in South Korea, the government proposed it as a solution for maintaining a reasonable size of the productive population and supporting the senior citizens to sustain their living. However, the younger generations tend to oppose the issue due to the fear of reduced opportunities for their jobs, as the plan would increase the labor cost for the companies.

The second issue is whether to classify game addiction as a disorder (Game Addiction) [126, 127]. The controversy on the second issue was initiated by the World Health Organization’s (WHO) decision to include gaming disorder in the International Classification of Diseases (ICD). However, considering the people’s negative perceptions of games and the Korean government’s previous attempts to regulate games, game players and the game industry perceived the Game Addiction issue as another example of imposing excessive regulation on games.

For the opinions in the system, we used data from the deployment study, as well as the data from two rounds of surveys on people’s opinions on the issues, which were conducted separately. Survey participants were recruited by a commercial online survey agency, with their age and gender equally balanced. The participants’ responses during the study were also collected and presented to other participants, yielding a total of 515 responses for the Retirement issue and 461 responses for the Game Addiction issue.

Figure 3.5 and 3.6 present how the opinions were distributed in the system. For the Retirement issue, a significant difference in opinion distribution was observed only by age groups (Kruskal-Wallis test, $p = 0.0002$). Post-hoc analysis with Dunn’s test with Bonferroni correction showed that the opinion distribution of age group 20-29 significantly differed from that of age group 30-39 ($p = 0.000144$). For the Game Addiction issue, a significant difference in opinion distribution was observed by age groups (Kruskal-Wallis test, $p < 1e-8$) and gender (Mann-Whitney U test, $p = 0.00002$). Post-hoc analysis with Dunn’s test with Bonferroni correction showed that the opinion distribution of age group 20-29 significantly differed from that of age group 40-49 ($p = 0.0002$), 50-59 ($p = 0.0006$), and 60-69 ($p < 1e-6$), and that of age group 30-39 significantly differed from that of age group 60-69 ($p = 0.0467$).

3.4.3 Procedure

For the first round of user study, we conducted 12 group-based usability testing sessions. Each session lasted for approximately 75 minutes, including pre-/post-study surveys. For the second round of user study, we conducted six individual sessions with a length of approximately 90 minutes. Every session was conducted remotely over Zoom. During the session, the researcher gave verbal instructions for the tasks, and the participants used the system with their computers.

First, the participants were briefed about the goal of the study and the overview of the task. Then the participants answered a pre-study survey. We asked for their prior knowledge, interests, and opinions on the two issues for the experiment. We also measured their argument repertoire on the issues and surveyed attitudes towards people with incoherent opinions on the two issues.

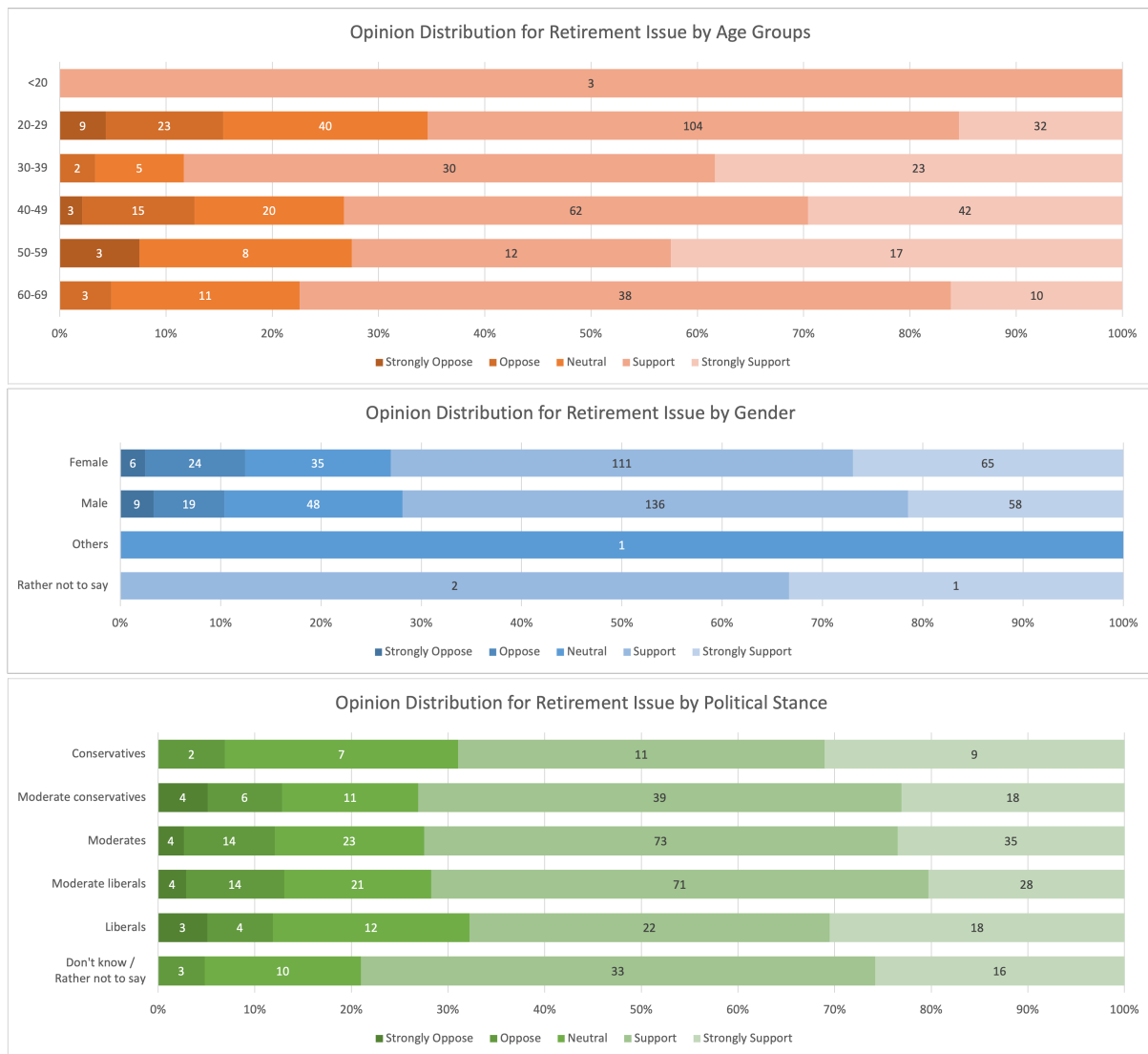


Figure 3.5: Distribution of opinions on the Retirement issue used for the study.

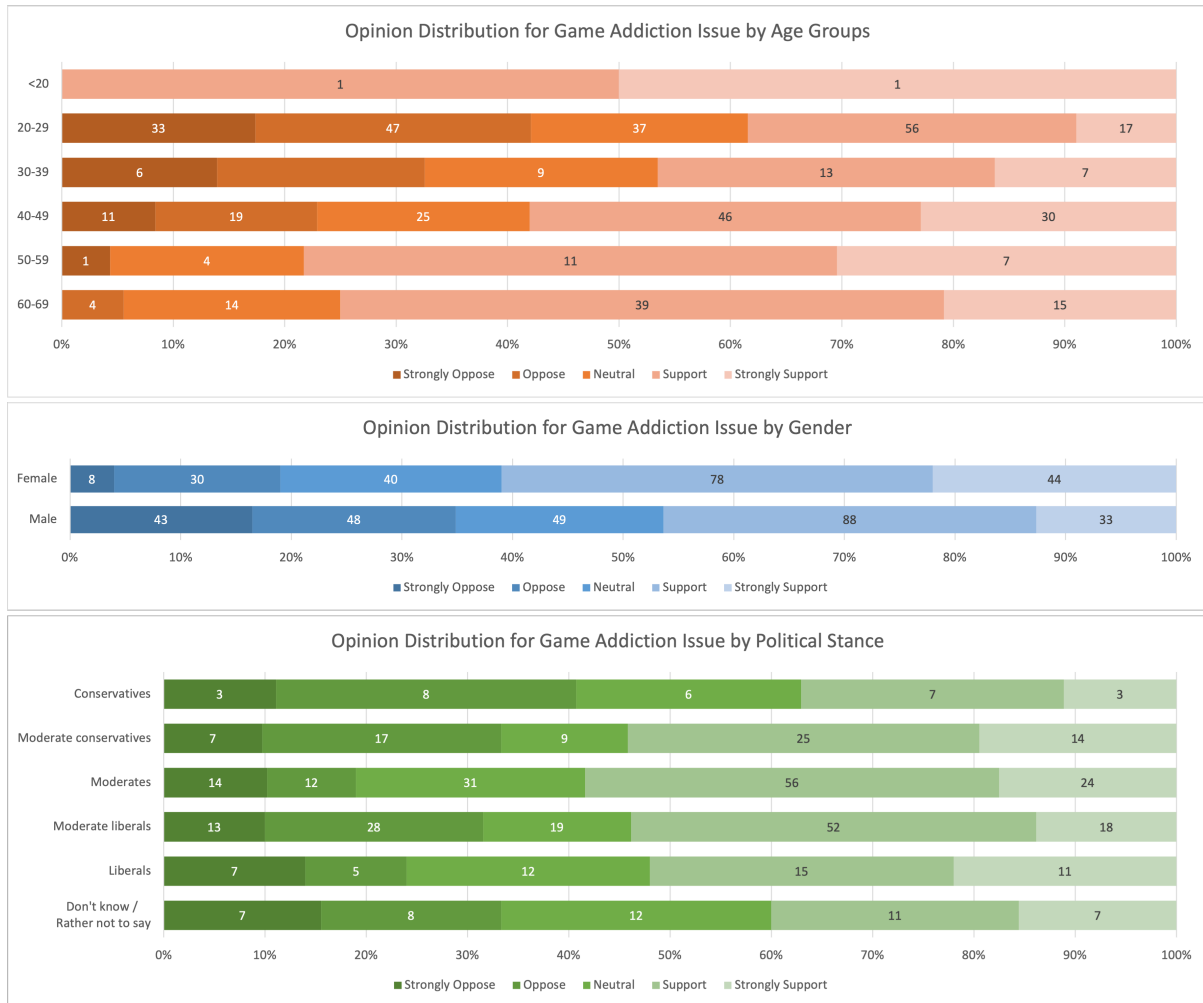


Figure 3.6: Distribution of opinions on the Game Addiction issue used for the study.

Next, they were asked to sign up to the system and leave their own opinions on the two issues.

After a brief overview of the system, the participants used the system to understand the overall opinion landscape on each issue. We did not impose any specific restrictions on using the system, other than asking them to use the system for more than five minutes to ensure that they had a thorough experience with the system. All participant activity, such as which dots they clicked or hovered on and which set of filters were active while reading opinions, was logged.

After using the system, the participants answered a post-survey about their experience with the system. We asked them if their opinions on the issues did or did not change after using the system, their reason for doing so, and their attitudes towards people of the opposite stance. We also measured the argument repertoire [128] on the issues after using StarryThoughts, as well as comments on the experiences and the overall level of satisfaction they felt from using the system. In the second round, we conducted user interviews instead of the survey, focusing on their experience of using demographic identities for exploring and reading opinions.

3.4.4 Measures

Experience of using StarryThoughts

To understand how the participants used our system, we analyzed the activity logs of the participants. When logging, we decided that the participant “read” an opinion if the opinion box was visible for more than 0.5 seconds. With such a tight time limit, we tried to capture every opinion the participant paid attention to. We also logged the characteristics of each opinion that the participants read, such as what kinds of filters were applied when reading the opinions and whether the opinion was a recommended opinion at the time of reading.

Opinion quality

To gauge how the quality of the opinion changed before and after using StarryThoughts, we adopted the argument repertoire [128], a list of supporting or opposing reasons related to an issue. The measure was used in deliberation research to measure the quality of participants’ opinions [128, 129]. To measure the argument repertoire, we asked the participants to list the reasons why supporters of the issue would support it and why dissenters would oppose it in the pre-/post-task survey.

To compare the argument repertoire before and after using the system, we sampled the responses of 30 participants and coded them with the help of two external evaluators. We thought it might not be likely that the participants forgot the reasoning after the task, so we evaluated in an additive manner as follows. First, we grouped the pre-task arguments and post-task arguments as a set, yielding four sets of arguments (Retirement / Game Addiction issues $\times$ Supporting / Opposing). The evaluators identified all relevant reasons to support or object to the issue from the pre-task responses. Then, they reviewed the post-task responses in comparison with pre-task responses and identified newly added arguments after using the system.

While identifying the arguments, we asked the evaluators to count only the arguments relevant to the issue (i. e., not a mere preference or emotional response to the issue). Also, we did not consider the factual accuracy of the arguments. We followed the practice of the original paper, which measured the argument repertoire [128], which justified the practice as it would be tough to verify the reasons and evidence supporting the claims.

Table 3.1: Perceived importance and prior knowledge level of the issues used in the study. Results suggest that the participants considered the Retirement issue as more important but knew less about it. (Wilcoxon signed-rank test, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$)

		Retirement	Game Addiction
First round	Perceived importance ***	5.30 / 7	3.82 / 7
	Prior knowledge **	3.46 / 7	4.46 / 7
Second round	Perceived importance	5.17 / 7	4.00 / 7
	Prior knowledge	3.67 / 7	4.67 / 7

To build a consensus on the coding criteria, the evaluators worked on a set of supportive arguments for the Retirement issue together. With consensus on the rating process from the collaborative rating, the evaluators individually evaluated the remaining three sets of arguments and resolved the conflicts afterward. Due to the complex nature of the coding process, we could not directly report the inter-rater reliability. Still, the two evaluators’ results exactly matched for 63% of the responses that were individually evaluated.

Attitude towards diverse opinions

To gauge participants’ attitudes towards different opinions, we first observed whether there were any changes in opinion or stance after using the system. We used 7-point Likert scale to rate the quality of opinions in the system, including whether the opinions were logical, well-supported, agreeable, provided new perspectives, and helpful for deciding their own opinion. With the open-ended responses on the opinion changes and the ratings on the opinion qualities in the system, we aimed to understand how being exposed to diverse opinions could change their opinions.

As a more direct measure, we adopted the Respectfulness Composite survey questions from previous work on openness towards opposing views [130]. The composite consisted of four statements on why dissenters would disagree with the participant: “They come from a different cultural (e.g., racial, political, financial, religious) background than I come from”, “They do not understand the issue as well as I understand it”, “The issue is complex and warrants different opinions”, and “They are not as intelligent as I am”. The participants were asked to rate each statement on a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree). The ratings were averaged to yield a composite measure, with the ratings for disrespectful statements reversed.

3.5 Results

In this section, we present how participants used StarryThoughts and how the experience of using StarryThoughts affected the participants’ opinions towards social issues.

We first present participants’ perceived importance and prior knowledge level of the issues used in the study (Table 3.1). From the difference, we provide some clues on the different experiences of using the system, depending on the issue characteristics. Overall, participants reported that they considered the Retirement issue more important than the Game Addiction issue. On the other hand, they responded that they knew less about the Retirement issue than the Game Addiction issue.

Table 3.2: Characteristics of opinions read by the participants from the first round of evaluation. The numbers suggest that people were more interested in the opinions of their own groups.

	Opinion		Age		Gender		Political stance	
	Same	Different	Same	Different	Same	Different	Same	Different
Retirement	45.0	24.1	32.2	36.9	38.3	30.8	36.5	32.6
Game Addiction	30.8	25.6	26.7	29.6	30.1	26.2	28.2	28.1

3.5.1 Strategies and patterns of exploration

To answer our **RQ1**, we analyzed participants’ system usage logs and their open-ended responses regarding their experience with StarryThoughts. In this section, we first present statistics showing the diversity of opinions browsed by the users. Then, we examine behavioral patterns and survey and interview responses to analyze how participants valued the system features.

Balanced opinion exploration

To gauge the extent of the diversity of opinions participants explored, we counted the number of opinions that the participants read. We defined the opinion as having the same stance as the participant if both supported or opposed the issue. If the participant was neutral on the issue, we defined neutral opinions, as well as moderately supportive or opposing opinions, as opinions with the same stance. Every opinion that did not align with the participant’s stance was counted as different. For age group and gender, we counted only the opinions from the participant’s group. We analyzed the opinion diversity with respect to the political alignment as we did in terms of the stance on the issue.

Table 3.2 shows the characteristics of opinions read by the participants. The results suggest that the participants were interested in the opinions of groups with shared characteristics. For instance, Q4 mentioned that he was interested in his peers’ opinions because he had to live with them, so he wanted to better understand their perspectives.

Exploring opinions with visualization

Participants benefited from the 2-D visualization, which enabled them to investigate the overall trend in opinions. Among 5,143 events of opinions being read, 3,341 (65.0%) events occurred without any filters on demographic factors. Participants mentioned that they tried to explore the opinions in a balanced manner in terms of demographic profiles as well as the stance on the issue. Quotes of Q6, who focused on reading the most extreme opinions first, revealed some underlying intent behind such behavior. He mentioned that he was curious about “how people with extreme opinions think”, while he thought “moderate people would have opinions that everyone would know already”. In other cases, the participants mentioned that they were selecting the opinions without specific criteria in mind.

Exploring opinions with identity factors

With the visualization of opinion distribution and filtering function, one of the most common usage patterns was to inspect the opinion trends of each demographic group. The participants observed the trend of opinions by trying the filter with every possible identity option and glancing at how the dots

were horizontally distributed across the 2-D space. Such exploration helped them quickly understand how the opinions on the social issues differed by demographic factors.

Participants used the filters to focus on opinions from a specific group of interest. Participants were commonly interested in the opinions from the groups they belonged to, as well as the expected supporters and dissenters. Participants were also intrigued by the groups that had unexpected trends in opinions.

Patterns of opinion distribution for each demographic group served as a starting point of exploration by separating the majority opinions from the minority ones. Doing so allowed them to read both types of opinions to learn how the selected groups think about the issue. For example, Q5 said, “I read the opinions when I looked into a new group to understand why each group thinks the way they do. (...) I chose 2-3 opinions from the majority opinion and 1-2 opinions from the minority opinions.”

When exploring a demographic group’s opinions, participants mentioned that the absence of a visible trend for each demographic group made the experience less engaging. For example, P37 mentioned that the absence of a clear trend made them feel frustrated, as they could not find a starting point for exploration. Q3, Q4, and Q6 suggested employing more apparent visual elements for visualizing the trend of opinions, such as presenting the histogram or directly showing the proportion. They expected to quickly compare the opinion trends between multiple social groups and recognize the differences in such elements. Q3 mentioned that custom categories more closely related to the issue could make the exploration more intriguing, for example, including their prior experience with games for the Game issue.

Usage of recommendation

Recommending different opinions for fostering diverse exploration of opinions was one of the main features of the system. The activity logs showed that among the total 5,143 events of reading opinions, 1,627 (31.6%) opinions were from recommendations. 689 of them were recommended as “Same Stance, Different Point of View”, 476 of them were recommended as “Opposite Stance, Different Point of View”, and 462 of them were recommended as “Opposite Stance, Similar Point of View”.

Out of 50 participants, 43 of them were favorable towards the recommendation of different perspectives. Participants mentioned that the recommendation helped them easily select different opinions to read and compare them against their own in an engaging manner. Participants especially liked the visual design of the recommendation of emphasizing the dots with glows. P11 said that its visual highlights motivated them to click the recommended opinions across the dispersed dots, rather than narrowly exploring a couple of dots near their own.

However, the participants shared some concerns about the recommendation. First, some participants felt the recommendation results did not agree with the description. From the first round of the study, four participants questioned the accuracy of the recommendation algorithm. P37 said, “I did not feel the ‘different points of view’ were really that different from mine”.

Another common concern was about the design of the algorithm itself. After discovering that the recommendation algorithm recommended minority opinions, Q3 and Q4 worried that the algorithm might not accurately show the representative arguments from each group. As an alternative, they suggested focusing on identifying representative opinions for each stance based on user ratings to present the summary of opinions accurately.

Table 3.3: Argument repertoire of the participants before and after using the system. The participants were divided into three groups according to their prior opinion on the issue: supporters, neutrals, and dissenters. The table shows the number of reasons supporting and disagreeing with participants’ own stance. For the neutral group, we reported the number of reasons that supported the issue in the column “Reasons for own opinion”. The numbers show that the argument repertoire of the participants increased after using the system. (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$)

		Reasons for own opinion		Reasons for disagreeing opinion	
		Pre	Post	Pre	Post
Retirement	Supporters (N = 20)	1.95	3.2 ***	1.35	2.65 ***
	Neutral (N = 6)	1.33	2.67 *	1.83	2.67
	Dissenters (N = 4)	1.5	2.25	1.5	3.75
Game Addiction	Supporters (N = 14)	1.64	2.79 **	1.43	3.36 ***
	Neutral (N = 4)	1.32	2.51	1.35	3.13
	Dissenters (N = 12)	1.17	2.58 **	0.83	2.08 **

3.5.2 Effect of using StarryThoughts for opinion strength and quality

To answer our **RQ2**, we analyzed how the participants changed their opinion after using StarryThoughts and how their confidence in the opinion changed. We included the quotes from the participants that could further explain the effect of StarryThoughts on their opinions.

Argument repertoire

The analysis result of argument repertoire showed that the participants had expanded their argument repertoire for both issues (Table 3.3). The participants were divided into three groups according to their prior opinion on the issue: supporters, neutrals, and dissenters. We analyzed the number of reasons consistent/inconsistent with their opinion with the Wilcoxon signed-rank test for each group with more than five participants. In all cases except for dissenters and neutrals for the Retirement issues, the participants could come up with at least one more reason for both supporting and objecting to their prior opinion.

Effect of using StarryThoughts on opinion strength

Tables 3.4 and 3.5 show how the participants’ opinions changed on the issues. The results show that the overall trend of opinions on the issues shifted towards less support for the issue, and the number of people with extreme opinions has decreased.

Exploring diverse opinions with the system helped the participants be more confident in their opinions on the issues. On the Retirement Age issue, the average score for the participants’ self-reported level of confidence in their own opinion increased significantly, from 4.18 / 7 to 5.0 / 7 (Wilcoxon signed-rank test, $p = 0.004$). Out of the 50 participants, 29 answered that their level of confidence increased for that issue, while 11 answered that they became less confident. For the Game Addiction issue, the average score of self-reported level of confidence in participants’ opinions also saw a significant increase from 4.38 / 7 to 5.02 / 7 (Wilcoxon signed-rank test, $p = 0.002$). For the Game Addiction issue, 25 out of 50

Table 3.4: Distribution of opinions on the retirement issue from the participants before and after using the system. While more than half of the participants kept their opinion, it could be observed that some of the participants with strong prior opinions became moderate after using the system.

Opinion on the Retirement issue		Opinion after using StarryThoughts				
		Strongly Oppose	Oppose	Neutral	Support	Strongly Support
Opinion before using StarryThoughts	Strongly Oppose	1	0	1	0	0
	Oppose	1	3	1	0	0
	Neutral	0	3	4	1	0
	Support	0	2	4	18	3
	Strongly Support	0	1	1	5	1

Table 3.5: Distribution of opinions on the Game Addiction issue from the participants before and after using StarryThoughts. The results show that some of the participants with strong prior opinions became moderate after using the system.

Opinion on the Game Addiction issue		Opinion after using StarryThoughts				
		Strongly Oppose	Oppose	Neutral	Support	Strongly Support
Opinion before using StarryThoughts	Strongly Oppose	7	4	0	0	0
	Oppose	1	8	1	0	0
	Neutral	0	5	3	1	0
	Support	0	5	3	10	0
	Strongly Support	0	1	1	0	0

participants answered that they became more confident, while 7 participants answered that they became less confident.

Open-ended responses from the participants showed some insights on how reading others' opinions could affect participants' opinions.

First, participants could reinforce their opinion by learning that there were others who shared similar opinions. P25 strengthened their opinion on objecting to the Retirement issue, saying, "I thought I was in a minor group, so I expected few people would share my opinion. (After using the system) I felt my opinion was more reasonable than I expected."

Second, the participants could discover new supporting arguments for their opinion from reading others' opinions. P31 said they strengthened their opinion on objecting to the Game Addiction issue after using the system. They said, "Among the opinions I read from the system, I found one questioning the specific criteria to define game addiction. I agreed with that opinion a lot, and I became more confident in my opinion."

Third, participants strengthened their own opinion after browsing incoherent opinions and realizing that they were not persuasive enough. Q4 mentioned that his opinion on dissenting from the Game Addiction became stronger after reading the supporting opinions. He said "I was surprised to see that people were supportive of the regulation on gaming with such shallow opinions. (...) I could understand them in the sense of personal experiences, but I could not agree with their opinion on how the social regulations or laws should be made."

On the other hand, some participants learned perspectives they had not considered from incoherent opinions and therefore became more prudent in their opinions.

For example, P32, who was initially supportive of increasing the retirement age, changed their opinion to neutral "after reading an opposing opinion pointing that older employees could be less passionate and cause a loss of efficiency in the workforce." P27 said they changed their opinion on the Game Addiction issue from opposing to neutral, saying "I saw an opinion from a person who suffered from game addiction, and they mentioned that external help was essential. Also, I agreed with the opinion that classifying game addiction as a disorder would help the actual game addicts to get good care."

3.5.3 Effect of StarryThoughts on evaluation of other's opinions

To answer **RQ3**, we analyzed how participants valued opinions in terms of respect and perceived quality.

Respect towards different opinions

The survey results show that the experience of using the system did not significantly affect the level of respect towards other opinions. Regarding the Retirement issue, the Respectfulness Composite index increased from 5.60/7 to 5.70/7 (Wilcoxon signed-rank test, $p = 0.055$). On the Game Addiction issue, the Respectfulness Composite index increased from 5.47 / 7 to 5.58 / 7 (Wilcoxon signed-rank test, $p = 0.133$). There were no significant differences in the component ratings of the composite index either.

Quality of other opinions

Figure 3.7 shows the score distribution of the participants' quality evaluation of the opinions they saw. We statistically analyzed the difference among five questions with the Friedman test and conducted post-hoc analyses using the Dunn test with Holm correction.

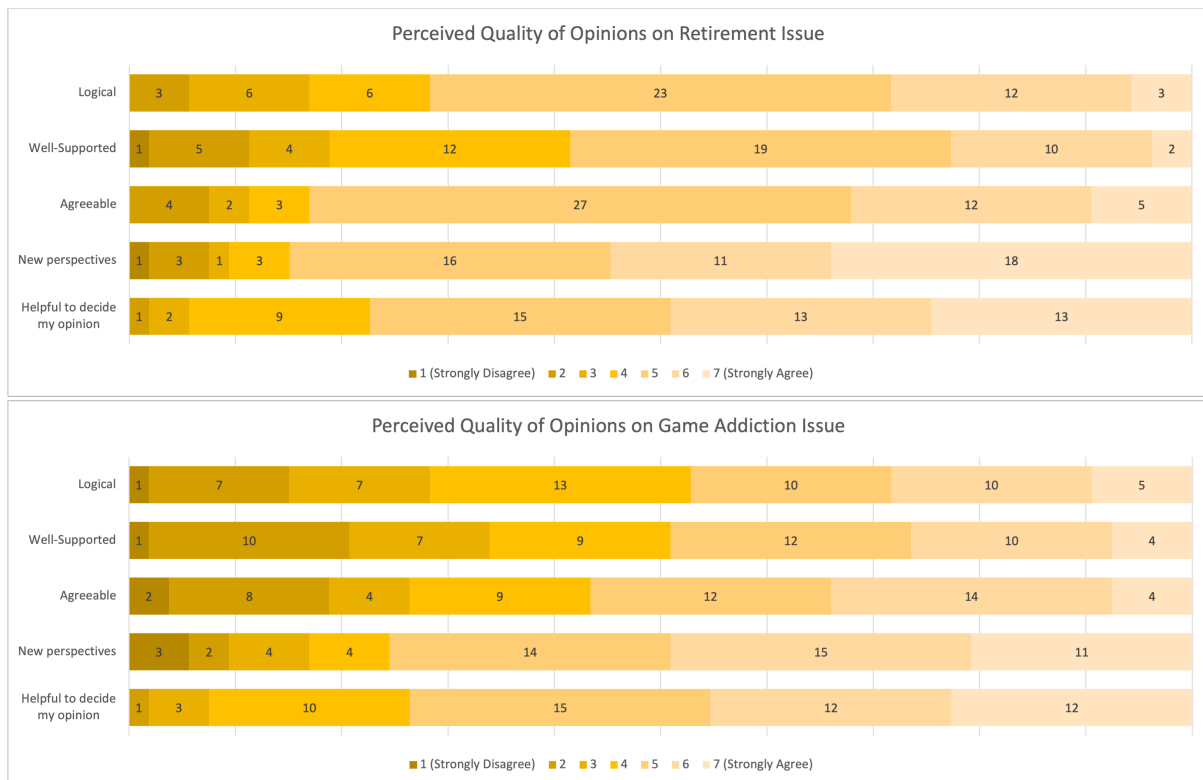


Figure 3.7: Distribution of participants' rating on the quality of the opinions used for the study. While participants were favorable to the opinions in general, participants felt opinions on Game Addiction issues were less agreeable and less novel.

While the participants felt generally positive about the opinion quality, the results showed that participants disagreed less that the opinions were logical or well-supported. For the Retirement issue, Logical vs. New perspectives ($p \leq 0.01$), Well-supported vs. New perspectives ($p \leq 0.001$), and Well-supported vs. Helpful for my opinion ($p \leq 0.01$) showed significant differences. For the Game Addiction issue, Logical vs. Helpful ($p \leq 0.05$) and Well-supported vs. Helpful for my opinion ($p \leq 0.05$) showed significant differences.

Participants commonly mentioned that there were a lot of arguments based on personal experiences. This had varying effects on the persuasiveness of an argument, mostly based on whether the participants did or did not have conflicting experiences themselves. In the case of the Retirement issue, participants mostly reported that they did not know of the experiences mentioned in the opinions, which had a positive effect on the perception of these opinions. However, in the case of the Game Addiction issue, participants were more skeptical towards opinions based on personal experiences because they were “different from what I have experienced.” (Q5) Such a difference is also visible from the difference in ratings between the two issues. The participants rated opinions on the Game Addiction issue as significantly less agreeable (Wilcoxon signed-rank test, $p = 0.037$) and less new (Wilcoxon signed-rank test, $p = 0.011$).

Value of identity factors

Participants mentioned that the demographic identities helped them understand the opinions better. Out of the 50 participants from the first round of evaluation, 44 responded that the feature was helpful.

One of the common benefits was having the opportunity to observe how different social groups thought about the issue. Q1 said, “I feel like people tend to flock together with people with similar opinions in online space. With the system, I could see the collected opinions of the whole public and filter them. I could acknowledge and resolve my misunderstanding towards social groups that I don’t belong to, and people do not share the same opinion with me.”

Demographic identities also positively affected the perceived credibility of the opinions. Q4 valued that the demographic identities made the opinions more credible compared to comments on online news, which presents “nothing about who wrote such comments.”

Participants also mentioned that they could empathize with the opinions of others when they saw their demographic profiles. Participants used the information on demographic identities to imagine what kind of interests or experiences the authors of the comments would have and how they would be related to the opinion. For example, Q4 said, “I read a supportive opinion on the Game Addiction issue, and I saw that the author is a woman in her 40s. I could understand why she ended up having such an opinion. Still, it’s separate from whether I agree with her or not.” However, this was not always the case, as in the example of Q2, who mentioned that for more politically controversial issues, he would try to see the opinions more skeptically if he knew the author’s political stance.

Some participants expressed worries that they might develop stereotypes or biases from associating opinions with individual identities. P18 said, “as I knew about the demographic identities of the author before actually reading the opinions, I realized that my prejudice was making me selectively accept the opinion.” Similarly, P7 said that they “felt like I had to reflect my bias in the process” and that “if we start perceiving these individuals as stakeholders, we may garner prejudices, thinking that they only think this because it affects them. That might hinder communication.” As a solution to this problem, P18 suggested that “maybe if the identities came after reading the opinion, or were placed below the text, it could have helped with the issue of bias.”

Value of expectation

Participants commonly mentioned that they would expect others' opinions to be aligned with their own interests or their experience related to the issue. It led to some common patterns in the participants' expectations of the trend of opinions. For example, on the Retirement issue, five out of six participants from the second round mentioned that people in their 20s and 30s would be the likely dissenters as they expected that the issue could shrink the job market. On the Game Addiction issue, all six participants expected that people in their 20s and 30s would be the likely dissenters because they expected that such a group would be more familiar with games.

Confirming expectations When participants discovered that the opinion was distributed as they expected, they considered that the groups were making rational decisions and respected such opinions. For example, Q4 shared his experience with the Retirement issue, "I could see what types of advantages or disadvantages each group would get from the issue so that I could understand both supporters' and dissenters' perspectives." However, with the Game Addiction issue, such experience worsened the negative stereotype of other demographic groups. For example, P2 shared their negative experience of reading opinions on the Game Addiction issues. They mentioned, "I couldn't help laughing as the supportive groups and dissenting groups were almost like what I expected. (...) I hope the supportive group could play games at least once before leaving opinions. I almost got mad at them because they were just saying 'Games are bad, it's a disease'".

Unexpected trends encourage understanding others On the other hand, when the participants discovered that the expected trend did not exist or even the opposite trend existed, participants were surprised and intrigued to explore why the social group actually thought differently.

From such exploration, the participants could widen their perspectives on the issue and understand other social groups better. For example, Q6 found out that the people in their 40s opposed the Retirement issue more than he expected. From their arguments, he learned that they were skeptical on the issue because of their negative personal experiences of working with older colleagues. P27 mentioned that they could widen their perspectives when they read the opinions against the initial expectation. From such opinions, they learned that people had considered multiple aspects of the issue, even if they had not aligned with the group's interests.

3.6 Discussion

In general, the evaluation results suggest that StarryThoughts could support the users to explore a wide range of opinions. In this section, we discuss some of the notable findings from the participants' behaviors and comments that would guide future research in this domain.

3.6.1 Role of identity

Participants' responses and activity patterns suggest that the identity factors supported exploration and interpretation of opinions in several dimensions. Demographic identities made the opinions more credible, and they helped participants understand the whole picture of the opinion landscape by demographic groups. Furthermore, participants used the demographic identities to guess the author's personal

Table 3.6: Comparison between the design of StarryThoughts and existing systems.

	ConsiderIt [53]	Opinion Space [51]	StarryThoughts
Opinion visualization	Pros/cons list	Dots on a 2-D space	Dots on a 2-D space
Visualization criteria	Stance on an individual topic	Similarity of opinion profile	Stance on an individual topic
Main user interaction	Building their own pros/cons list	Freely exploring opinions	Freely exploring opinions
Exploration strategies	Pros/cons list nudging for balanced consideration of the issue	Freely exploring opinions on a 2-D space	Freely exploring opinions on a 2-D space with demographic profiles
Recommendation criteria	Opinions appealing to diverse user groups	Opinions agreeable to diverse users	Opinions distant from the one being read

interests, experiences, and circumstances related to the issue. Based on such a prediction, the participants could interpret opinions in the context of the authors. When the actual opinion agreed with the expectation, participants could empathize with and respect the opinion, regardless of their stance. On the other hand, when participants discovered opinions were against their expectations, they were motivated to read the opinions and widen their perspectives on the issues and other social groups. However, participants also pointed out that their preconceptions or mental images of specific demographic groups could be a source of prejudice on other demographic groups.

Finding the right balance of using individual identities for attenuating prejudice would be the next item for future research. To reduce stereotypes, Richards and Hewstone [131] suggest that merely presenting instances against stereotypes would not be successful. The simple presentation would lead to subtyping, where such counterexamples are perceived as an exception to the existing stereotype, while not changing the stereotypical perception towards the group. To affect the stereotype, the perceiver should undergo a process of subgrouping. In subgrouping, the perceiver realizes that the stereotypical groups are not a single cohesive entity but composed of different subgroups with their characteristics [132, 120]. Through this process, individuals can gain subgroup salience, and in turn, their level of prejudice towards outgroups is reduced. This process could be implemented by introducing additional axes that could split the demographic groups into smaller subgroups.

3.6.2 Effect of visualizing opinions

Our system visualized the opinions of the users as dots on a 2-D space, similar to Opinion Space [51]. One difference between StarryThoughts and Opinion Space was how the system placed the opinions on a 2-D space (Table 3.6). Unlike Opinion Space, we explicitly used the stance of opinions to position them. Although we did not analyze the participants' usage patterns in a controlled experiment, the open-ended responses from participants suggested that the system was engaging and helpful for browsing diverse opinions on social issues. We observed that the participants easily understood the arrangement of opinions and used it to discover the distribution of the opinions. One participant also mentioned that

they could mentally prepare themselves to see opposing opinions by exploring the opinions by themselves.

Visualization of StarryThoughts allowed the participants to easily explore the opinions in a balanced manner by selecting opinions from extreme sides. This approach is based on values similar to ConsiderIt [53], which also supported the users to see more balanced aspects of the issue and discover comments that could help verify their own decision. In addition to this, StarryThoughts enables the users to discover a wider variety of arguments by not using a collaborative filtering approach in recommending the opinions. Such responses suggest that the “starry night” visualization could support people to accept diverse opinions.

When designing the system, we expected that the demographic identities of the authors and the distribution of opinions could enable the users to discover unexpected trends (or the lack thereof) of opinions for demographic groups or for them to encounter non-stereotypical opinions from a specific demographic group. By doing so, we expected that the users of StarryThoughts would be able to look at diverse opinions on social issues.

The participants’ actual usage pattern showed that they were able to follow such a flow. However, we also learned that discovering a clear trend of opinions was necessary for the participants, as discovering the opinion trend was the starting point for exploring the opinions.

We think an improved design of StarryThoughts would need to present not only the trend of the opinions in an objective scale, but the trend of the actual arguments for each demographic group. For example, topic analysis and visualization of arguments for each demographic group could be possible. For example, Baumer et al. [92] proposed a “words web” visualization of text for presenting the main ideas of a text while removing the framing effect from the text. Such visualization may contribute to the better delivery of the main points, so that the user can understand what ideas were used by each demographic group. It could be possible for the system to present what kind of social groups are dominant in the opposition for each user, so that the user could understand opposing opinions better.

3.6.3 Value of recommendation

While the recommendation feature helped the participants explore a diverse range of opinions, the evaluation result raises interesting questions about the design of algorithmic recommendations and the justification for using them.

Unlike previous work that used the appeal to diverse populations as the main criterion for recommendation [53, 51], our system used diversity as the main criterion for recommendation (Table 3.6). We observed that the participants were favorable to the recommendation feature and how it is visually represented. However, participants questioned how the recommendations were produced and whether they were actually representative opinions from each demographic group.

Although we implemented algorithmic recommendations as a way of facilitating exploration of diverse opinions, there were some limitations to our design. For example, our recommendation results were based on document embedding vectors computed by the weighted average of word vectors. This algorithm uses the inverse word frequency as the weight for computing the vectors. Therefore, it could be susceptible to opinions with rare words, which would cause them to be distant from all the other opinions in the vector space. The algorithm, in turn, could be biased to this specific subset of opinions. Furthermore, as diversity-oriented recommendations would treat every opinion equally, regardless of how much of the population agrees with it, users might end up developing a distorted image of the opinion landscape.

To prevent users from misunderstanding the opinion distribution, the system would need to present not only the results of the recommendation but also how they were obtained, and what kind of pitfalls might exist in the results.

Another interesting design consideration regards how to present the recommended opinions. StarryThoughts purposefully utilized simple visual cues - adding a ‘glow’ to each dot - as an indirect mode of suggestion, keeping the level of intervention to a minimum and allowing the user to acknowledge or ignore the suggested opinions freely. At the same time, the participants needed to select the recommended opinions by themselves, which may incur an additional mental and physical load for the participants. So, there were several participants demanding a feature that shows all recommendations at once. We also provided three different modes of recommendation that were all simultaneously provided to the user at the same level. Other methods of recommendation could include using more direct indicators, reducing the number of recommended options, limiting the types of recommendation modes being provided, or even introducing new modes of recommendation.

3.6.4 Generalization to other issues and cultures

In this chapter, we selected two controversial social issues in South Korea to evaluate the system. As participants’ prior knowledge and perceived importance of the issue differed in the two cases, the reactions were different as well. For the Retirement issue, participants had less prior knowledge on the issue, which might have caused the participants to be more open to others’ opinions. On the other hand, the participants had more knowledge of the Game Addiction issue, so they might have been more critical.

One important remaining question is how people’s perception of others’ opinions and use of the system would change for the issues with different characteristics.

While we selected controversial issues at the time of evaluation, the issues did not have a strong association with people’s political beliefs or identities. This may have caused the participants to take a more rational stance and be more expectant towards others’ responses. However, user reactions might differ if the issues were more sensitive and divisive by political beliefs or identity factors, such as gender conflicts or affirmative action.

In such cases, it might cause users to reinforce prior stereotypes or even develop new stereotypes from observing a polarized opinion distribution. One possible mitigation approach would be to cluster and subgroup the opinions by their argument similarities in addition to demographic identities. By doing so, users could discover arguments they could agree with and possibly break the stereotypes towards other groups. Issues related to minority groups could be an interesting domain to use demographic identities for opinion exploration since presenting explicit identity cues could help users discover opinions from such groups.

Another possible exploration would be to generalize the system to different cultural contexts. First, the list of critical identities may change according to the cultural background. Our system is built based on the social issues and demographic identities in South Korea, which is widely considered a predominantly monoethnic country. We expected that users would not be as interested in exploring opinions by ethnicity as they would be with age, gender, or political alignment. On the other hand, if the system were applied in a polyethnic culture such as the United States, the system would need to be modified to encompass ethnicity as an important identity factor of the authors. Second, cultural contexts may affect the users’ expectations of opinion distribution by demographic factors. These trends may differ depending on the cultural and historical context of each society. It is also possible that the same issue could be framed differently according to the diverse value systems of each society.

3.6.5 Limitations

The design features of StarryThoughts enable users to notice the variety of opinions of people regardless of their demographic identities. However, the benefit of using StarryThoughts was not fully explored in our study because the study participants were mostly young college students, as they were mainly recruited from university communities. Therefore, we could only understand the experience of exploring and evaluating others' opinions from the perspectives of the younger generation. Further studies with more diverse populations would help us understand how other demographic groups perceive the opinions through the system. Another limitation of our study is in the demographic distribution of the opinions in the system, which was skewed towards younger people in terms of age and towards more progressive people in terms of political stances. As a result, the experience of using our system might have been limited for the participants, as they might not see enough perspectives that are not from their own demographic group.

3.6.6 Future Work

A possible future research direction would be to investigate the effects of the spiral of silence [133]. From the current evaluation, we did not find any evidence that participants with minority opinions experienced threats from expressing minority opinions. Participants with minority opinions could learn more about the issues by exploring opposing opinions, or even gain more confidence in their opinions by learning that more people shared their opinions than expected. However, we think the threat of the spiral of silence could remain as the users are able to see the opinion distribution at a glance, both at a higher societal level and a lower social group level. We think a deployment study without pre-populated opinions targeting the general public, who are likely less interested and less open to diverse opinions compared to voluntary participants, could be the next step in investigating the issue.

One of the remaining challenges is to scale up the system to encompass the opinions of more users. With the current visualization, adding more opinions causes a denser display of dots, which could reduce usability when selecting and exploring the opinions. Interactive visualizations for presenting overviews and detailed pictures, such as Overview+Detail [134], as well as clustering opinions, could be a possible solution to expand StarryThoughts into a larger scale.

3.7 Conclusion

This chapter explored the idea of using the identity of individuals to discover diverse opinions around social issues and encourage respectful attitudes towards them. We built StarryThoughts, an online system for presenting the diversity of opinions and supporting exploration of them with identity-based filtering and content-based recommendation of opinions. Findings from the user study suggest that demographic identities could be a meaningful cue for exploring diverse opinions around social issues and gaining more respectful attitudes toward others' opinions. With StarryThoughts, we envision an online space where people can exchange opinions on social issues with respect and openness for those with different perspectives.

Chapter 4. DataDive: Supporting Readers’ Contextualization of Statistical Statements with Data Exploration

This chapter presents DataDive, an interactive system to support people in exploring more diverse information. This chapter has adapted, updated, and rewritten content from a paper published at ACM IUI 2024 [135]. All uses of “we”, “our”, and “us” in this chapter refer to the coauthors of the aforementioned paper.

4.1 Motivation & Contributions

Statistical information is commonly used to substantiate the magnitude of social issues. For example, data journalists produce news articles based on statistical data analyses, from economic indicators to scientific climate measurements [136]. During the COVID-19 pandemic, public health experts used statistical data to convey the severity of the pandemic, support the implementation of safety measures, and refute fake information [137, 138, 139]. Even laypeople produced and shared their own interpretations of statistical data in online forums to spark discussion on diverse issues [140]. Such texts often contain *statistical statements* (e.g. “Over the last 50 years, its [Sweden] fertility rate has fluctuated significantly, rising roughly in tandem with economic booms” (Figure 4.1), which describes the current state of the world with the support of statistical information.

The significance of data lies in its ability to derive meaning across various contexts, such as points of reference and correlated factors. To facilitate this interpretative process, readers should be able to posit statistical statements within diverse contexts and formulate personalized interpretations of the data [141]. In this chapter, we term this approach as “contextualization”. Consider a reader reading a news article claiming “In South Korea, the fertility rate — the average number of children born to a woman in her reproductive years — is now 0.78” [142]. The reader could ask questions to interrogate this statement such as comparing the fertility rate to other developed countries or the reader’s own country, or checking the temporal change in fertility rate to examine whether there exists a clear trend. Going deeper, they may formulate hypotheses to explain the phenomena, such as attributing the low fertility rate to a shortage of affordable housing, and exploring the correlation between the fertility rate and housing prices over the past decade.

Existing research on social data analysis [143] and comments on data-driven articles [140, 136, 144] has shown that readers contribute to the contextualization of data in the public discourse. However, readers often struggle to produce their own contextualization because the statistical statements only include the minimally necessary data to support the authors’ messages. Therefore, readers need to search for information to contextualize the statistical statements by themselves, which requires a complex cognitive process for the readers with barriers in each step [141, 145]. While existing work explored interactions for augmenting documents with data to provide greater contexts to statistical statements [63, 64, 65], their support is limited to providing data directly related to the text, thus confining the scope of context that readers can explore.

To support the exploration of relevant contexts beyond those directly related to the text, we propose DataDive, an interactive tool for contextualizing statistical statements by exploring relevant data while reading online text (Figure 4.1). With DataDive, the reader can explore various contextualizations by

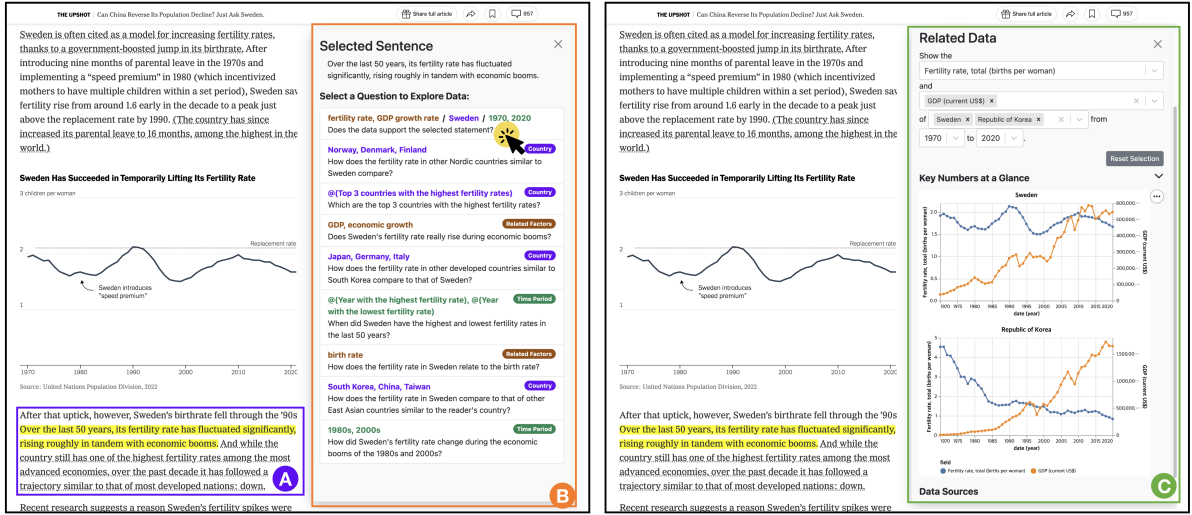


Figure 4.1: Overall user flow of DataDive. (A) While reading, the user selects a statistical statement of interest to start exploring contextualization. (B) On the side, DataDive generates and presents a set of recommendations suggesting different potential contexts around the statistical statement. (C) Upon selecting a recommendation, DataDive matches the most relevant data from its database and visualizes the data. The user can use dropdown widgets to explore data in different contexts.

picking a statistical statement of interest in the text and exploring system-generated recommendations of potential contexts to explore. When the reader selects a potential context, DataDive provides an interactive visualization of the selected contextualization of data, using external data from reliable sources. The reader can explore diverse data relevant to the selected contextualization by browsing the list of data series provided by DataDive. We built an LLM-based pipeline to parse statistical statements, which often contain vague and implicit references to data, for generating recommendations for contextualization and retrieving the most relevant data points from datasets containing a number of data series.

To verify if the pipeline of DataDive can generate recommendations for potentially meaningful contexts and retrieve relevant datasets for a given statistical statement, we conducted a technical evaluation with 18 external evaluators on 77 statistical statements on various topics around global social issues and a dataset of global statistical indicators. Our technical evaluation results showed that 96.6% of pipeline-generated recommendations were considered meaningful for contextualization, and the pipeline could retrieve relevant data for 80.5% of the statements. To evaluate whether DataDive can support readers in contextualizing the statistical statements from online texts, we conducted a within-subjects study with 21 participants. In the study, participants read two news articles with statistical information on global social issues while using either only a web search engine or a web search engine with DataDive to explore contextual information of interest. Findings from observations and interviews showed how DataDive supported the participants in exploring more contexts around statistical statements and in better contextualizing the statistical statements in the text, while we did not discover statistically significant differences in participants' knowledge gain from their reading process. We further discuss potential improvements for DataDive to support more diverse types of contextualization and design lessons from using LLMs to produce factually reliable yet exploratory results.

This chapter provides the following contributions:

- DataDive, a system for supporting contextualization of statistical statements while reading online

texts.

- A technical pipeline for generating recommendations for contexts and retrieving datasets relevant to statistical statements on the web.
- Findings from a technical evaluation demonstrating the effectiveness of our pipeline in identifying relevant contexts, and from a user evaluation showing that DataDive can support readers' contextualization of data while reading online texts.

4.2 Formative Study

We conducted a formative study to understand in what situations people seek related information to contextualize statistical statements in online texts. Based on the findings, we propose design goals for supporting the contextualization of statistical statements.

4.2.1 Study Setup

We recruited seven participants who read daily online discussions on social media or online forums. The call for participation was posted on an online board of a university in South Korea. The average age of participants was 23 (Min = 21, Max = 25). We screened the participants by asking how often they read and participate in online discussions.

In the study, we chose 'online discussions' as the context, as we expected that online discussions would often refer to statistical data in a less rigorous yet more argumentative manner, which would lead to more natural motivations to better understand the data. We provided four discussion threads with references to statistical information from Reddit's r/changemyview¹, which is well-known as a place for civil discussions [146]:

- We should reward those who are in great shape with yearly bonuses with a tax on junk food²
- It is unethical to purchase residential properties for investment purposes³
- The gender pay gap is largely caused by differences in goals in life and not by systematic discriminatory work practices⁴
- Nuclear Energy is way better than people think⁵

Participants were asked to choose two threads based on their interests and read them while thinking aloud for 20 minutes per thread, with the minimum requirement of reading the original post and three top-level comments. They were free to explore external information, such as clicking embedded links in the comments or searching for information on the web. After reading the threads, the participants were asked to search for external information of interest on the same two discussion threads for 10 minutes per thread. With a separate stage for searching for external information, we aimed to observe cases

¹<https://reddit.com/r/changemyview>

²https://www.reddit.com/r/changemyview/comments/13quoqw/cmv_we_should_reward_those_who_are_in_great_shape/

³https://www.reddit.com/r/changemyview/comments/12cqfm0/cmv_it_is_unethical_to_purchase_residential

⁴https://www.reddit.com/r/changemyview/comments/w4kbnv/cmv_the_gender_pay_gap_is_largely_caused_by/

⁵https://www.reddit.com/r/changemyview/comments/ud5hde/cmv_nuclear_energy_is_way_better_than_people_think/

where participants had latent informational needs but did not spontaneously search for them. Next, we conducted a semi-structured interview with each participant on their information-searching behaviors and current challenges. Each session lasted 90-120 minutes, and participants were compensated KRW 20,000 ($\approx$ USD 15). After all sessions, we analyzed our observation notes and interview quotes with affinity diagramming.

4.2.2 Design Goals

From the interviews and observations, we discovered participants' common behaviors and challenges in finding information to contextualize and understand statistical information. Based on them, we propose three design goals for an interactive system to contextualize statistical statements from online texts with data:

G1. Provide Straightforward Responses to Users' Information Needs with Data. Participants (P1, P5, P7) commonly considered reading as the main goal, therefore expecting only a low cognitive burden for the secondary goal of seeking information. They often wanted to quickly obtain answers to their questions without disrupting their focus on reading discussion threads.

Once they decided to search, participants needed to formulate a search query expressing their information needs. However, when they were unsure about what type and scope of information was appropriate, they felt it was challenging to start information exploration (P3, P5, P6). Therefore, we propose that the system be able to intuitively support users' informational needs while minimally disturbing the reading process.

G2. Facilitate Flexible Exploration of Contextual Information Around Data. While participants commonly used quick answer panels from the search engine or top-ranked documents to locate the data of interest, some participants (P1, P5, P6, P7) often wanted to seek data beyond what was available from a single source, such as comparing multiple data points or even different statistical datasets. Such cases were more common when they were verifying comparative statements from the discussion or wanted to explore contexts involving personal relationships or curiosity. However, retrieving and combining multiple pieces of data required excessive effort of searching multiple times and cognitively processing them, making it challenging to conduct such comparisons by themselves. In such cases, the system should provide relevant data based on the user's diverse information needs and support their exploration of the context surrounding the issue and themselves.

G3. Enable Users to Evaluate the Reliability of Provided Statistical Data. Among the search results, participants sought relevant and credible information that answered their questions. They typically tried to verify suspicious claims within online discussions. Participants often evaluated the reputation of the websites providing the information (P1, P2, P4, P6, P7). However, they rarely looked deeper into the data itself, such as the source of the data and the definition of measurement, to assess the credibility of the data and its relevance to the users' information needs. We propose that the system should provide information on the data to assess its credibility and relevance, such as the producer of the data, the definition of the data, and the link to the raw data.

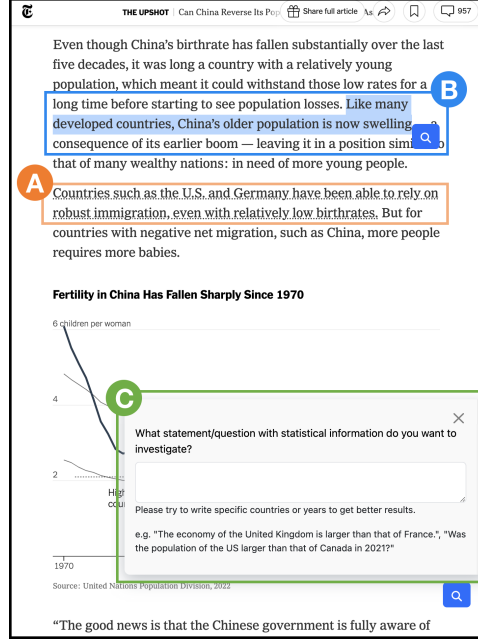


Figure 4.2: Three ways to start exploring contexts in DataDive. The user can (A) click on underlined statistical statements, (B) select the part of the text by themselves, or (C) type in their own statement of interest.

4.3 User Interaction

We present DataDive, a system for facilitating the contextualization of statistical statements with data, and recommending questions to guide contextualization and interactions for exploration of statistical data surrounding the context of the text being read. DataDive is developed as a browser extension, and we also built a standalone version for the controlled experiment. With DataDive, the user can first specify the statistical statement of interest. Then, DataDive presents a list of recommendations for contextualization, which the user can select to explore the data with interactive visualization.

4.3.1 Step 1. Initiating Context Exploration

To support exploring data with low cognitive burden (G1), DataDive provides three ways for users to start data exploration and express their needs for diverse types of data (Figure 4.2). First, the user can click on an underlined sentence within the text (Figure 4.2A). To support users in noticing statistical statements and facilitating data exploration, DataDive pre-processes the text to mark out sentences with statistical information. Second, the user can select any text snippet and click the search button if they want to investigate a specific part of the text (Figure 4.2B). Third, the user can formulate their own statement of interest with a free-form input (Figure 4.2C).

4.3.2 Step 2. Browsing Recommended Contexts

Upon selecting or entering a statistical statement, DataDive presents a list of questions to guide the exploration of diverse aspects around the statistical statement (Figure 4.3), supporting the exploration of contextualization with less burden (G1). The first option is fixed to show the most relevant data for the provided statement (Figure 4.3A-1). Then, DataDive presents a list of pipeline-generated



Figure 4.3: (A) UI for browsing the list of recommended questions and exploring data. (1) The first question is to verify the sentence with data. Each recommendation consists of (2) a question suggesting a different context and (3) the values of statistical indicators/entities/time periods. (B, C, D) After selecting one of the question from (A), DataDive provides (B) dropdown widgets to change selected statistical entities, indicators, and time periods, (C) visualization of selected data, and (D) metadata on data definition and source information.

Table 4.1: The rules of DataDive for generating visualization.

Statistical Entities	Selected Indicators	Type of Selected Entities	
		Single Point	Duration
Single	Single	Bar	Single Line
	Multiple	Bar	Multi Line
Multiple	Single	Bar	Multi Line
	Multiple	Grouped Bar	Multi Line / Entity

recommendations for contextualizing the statement considering the readers' context, which is self-stated to the system when they first use DataDive (Section 4.4.2). Each recommendation consists of a question (Figure 4.3A-2) and the associated values of relevant statistical indicators, entities, or time periods (Figure 4.3A-3) to support users in expecting specific data they would see for each recommendation.

4.3.3 Step 3. Investigating and Exploring Data

When the user selects a recommended question, DataDive queries the pipeline to retrieve the relevant set of statistical entities, indicators, and time (Section 4.4.3), fetches the data values, and visualizes them. The data presentation panel consists of dropdown menus, data visualization, and data sources (Figure 4.3B, C, D).

To support users in flexibly exploring different statistical entities, indicators, and time (G2) as well as correcting the potential errors of the pipeline by themselves, DataDive provides dropdowns (Figure 4.3B). Based on the user's selection for each field, DataDive generates either a bar chart or a line chart to present the selected data based on a pre-defined rule considering the number of selected values per field (Table 4.1). We focused on bar charts and line charts as they are the most common types of charts [147]. For visualizing data, we used the Vega-Lite [148] library.

Existing work on data science and data journalism points out that the provenance of the data is

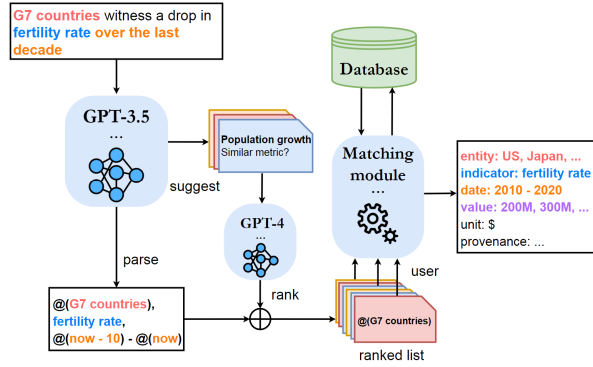


Figure 4.4: The pipeline starts by parsing a sentence into a tuple (entity, indicator, date) and generating related recommendations. These recommendations are ranked and displayed according to the ranks. User selection of a recommendation triggers a match with relevant data fields in top-k datasets. Finally, the system filters and presents the pertinent data.

important information for judging the credibility and the potential bias of data [149, 150, 151], which is aligned with our design goals (G3). DataDive provides a “Data Source” panel to show the metadata of the selected statistical indicators (Figure 4.3D), including the data definition and source information.

4.4 Technical Pipeline

DataDive utilizes a two-stage pipeline to recommend and retrieve contexts relevant to a statistical statement (Figure 4.4)⁶.

Our pipeline takes a statistical statement and its surrounding text as input. Based on a context database of datasets that can be used to generate contextual information, DataDive (1) generates a ranked list of candidate contexts from the input statistical statement (Section 4.4.2) and (2) matches the top contexts in the ranked list with the relevant data in the context database (Section 4.4.3).

4.4.1 The Context Database

DataDive first requires a database of datasets to map the statistical statements on. To support easier expansion of datasets, the pipeline takes each dataset as two CSV files. The first file contains the numbers, consisting of one column containing the names of entities, one column containing the date information, and the other columns containing the values per each indicator, following a common structure used in previous work [152]. This format simplifies the integration of new datasets, providing a consistent and general structure for the data. The second file contains the metadata about each indicator/feature in the dataset that would be presented to the users, including the unit, the source of the data, and the explanation of the data.

For the evaluation, we prepopulated the database with a variety of datasets on social issues from reputable sources: greenhouse gas emissions and energy consumption data from Our World in Data [153, 154], and the World Development Indicator datasets from the World Bank [155], which encompasses a wide range of topics on global issues (e.g., global health, economics, education, environment, technology).

⁶The code for the pipeline is available at <https://github.com/kixlab/ClaimVis>

4.4.2 Stage 1. Generation of Ranked Candidate Contexts

In the first stage, our pipeline generates a ranked set of candidate contexts through a three-step process: (1) parsing the components of the statistical statement from the input sentence, (2) generating candidate contexts based on the parsed information, and (3) ranking the candidate contexts.

Step 1. Parse Statistical Statement in Input Sentence

First, DataDive parses the input sentence to extract the three key components of a statistical statement (Figure 4.5): (1) *(statistical) entity*, the subject of the statement (e.g., ‘Korea’); (2) the *feature* (e.g., value, trend) and the *indicator* used (e.g., ‘fertility rate’); and (3) *date*, the time point or period of interest (e.g., ‘2019’). While existing work tackled similar challenges of disambiguating

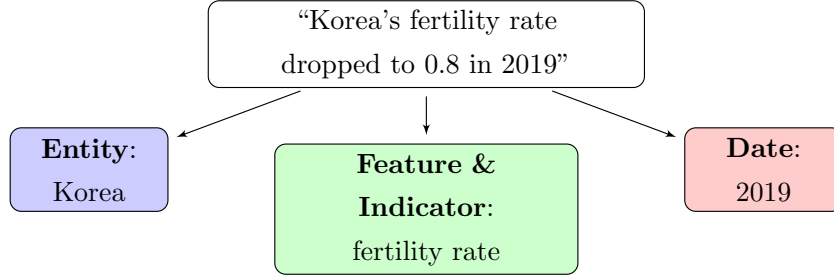


Figure 4.5: Example of parsing an input sentence into statistical entity, feature and indicator, and date.

natural language queries for data analysis [156, 157], our pipeline faces some unique challenges in handling diverse statistical statements from online texts which often implicitly refer to statistical data.

- **Feature & Indicator:** The value to be extracted might not exactly match the attributes in the tables or is only implicitly inferred from the statement. (e.g. “the number of people” implying “population”, “Korean woman gave birth to 0.8 children on average in 2019” implying “fertility rate”).
- **Entity:** The statement might refer to a group of entities that cannot be directly assigned to the names of entities (e.g., “Asian countries” or “Countries with a population of over 100 million”).
- **Date:** The dates in the statement might rely on implicit inference within the context of the whole text. (e.g. “the last decade” or “10 years after”)

Therefore, compared to existing work using semantic similarities with embedding vectors for disambiguation, we decided to leverage the capability of the LLM to handle such implicit references. In this step, such implicit references are first automatically annotated with a fine-tuned GPT-3.5 model, and are resolved in the later part of the pipeline (Section 4.4.3).

Step 2. Generate Candidate Contexts

Our pipeline generates a list of potential contexts based on the components extracted in Step 1. DataDive specifically considers four axes (Table 4.2): (1) *in-text*, which are based on the surrounding sentences; (2) *relational*, which are based on the similarities and relationships (e.g. causal, compositional); (3) *statistical*, which are statistically significant features along the component of interest (in our work, we specifically focus on global extrema); and (4) *personalized*, which are components that a reader is likely

Table 4.2: Criteria for generating candidate contexts.

Component	Candidate Context Generation Criteria
Entity	[In-text] Statistical entities referred to in surrounding sentences
	[Relational] Similar or related statistical entities
	[Statistical] Statistical entities displaying extrema
	[Personalized] Statistical entities the reader is likely familiar with
Features & Indicators	[In-text] Features & indicators referred in surrounding sentences
	[Relational] Similar or related (e.g., compositional, causal) features & indicators
	[Statistical] Global extrema features
Date	[In-text] Dates & time periods referred to in surrounding sentences
	[Relational] Dates & time periods with similar or related events (e.g., financial crisis)
	[Personalized] Dates & periods the reader is likely familiar with

to have prior knowledge about due to personal connections or media coverage that can help internalize contents of the statistical statement.

DataDive uses the GPT-3.5 model to generate the candidates of features/indicators, statistical entities, and dates based on the sentence of interest, its surrounding paragraph, and the readers’ personal contexts. Our pipeline also generates a teaser question for each context, as shown in Section 4.3.2. We recombine the generated candidates and the original components to generate the (entity, feature, date) tuples that we use as candidate contexts.

Step 3. Rank Candidate Contexts

Next, DataDive ranks the candidate contexts by these four criteria: (1) helpfulness in the broader context of the statistical statement, (2) interestingness to the reader, (3) novelty of the perspective, and (4) likelihood of discovering supporting (or refuting) data. Following a practice from a previous work [158], we used the GPT-4 model to run pairwise comparisons between each context, considering the four criteria, and combined the comparison results to rank the contexts.

4.4.3 Stage 2. Extraction of Relevant Data

When the user selects one of the candidate contexts from Stage 1, DataDive retrieves the relevant datasets in our context database and extracts the relevant data from the selected datasets through a two-step process.

Step 1. Retrieve Relevant Datasets

DataDive begins by retrieving relevant datasets from the context database to narrow the search space for relevant indicators. To do so, we compute the semantic similarity of indicators in each dataset and the components extracted from the statistical statement of interest.

The pipeline uses *all-MiniLM-L6-v2* Sentence-BERT model [159] to produce sentence embedding vectors. For the given context set, the pipeline computes embedding vectors for three components extracted from the statistical statement (Section 4.4.2), along with the values for the ranked candidate contexts. The embedding vectors are collectively used as a representative set for the original statement,

which we will refer to as the *statement set*. Similarly, the pipeline gathers *dataset set* for each dataset, consisting of sentence embedding vectors for each indicator name. To improve computational efficiency, these *dataset sets* are pre-computed and cached.

Our pipeline selects the top 7 datasets (empirically set) with the highest maximum attribute similarity scores. To reduce the dataset size for later steps, we only keep the top 15 attributes per dataset (empirically set) and merge the datasets into a single data table.

Step 2. Locate Context in Data Table

When the user selects a candidate context or the original context from the recommendation panel, DataDive locates the context within the data table. The pipeline then specifically tries to match the (entity, feature & indicator, date) tuples with the data table from Step 1. It first attempts to match each component directly with the table attributes. When the components do not match the attributes exactly, such as variations (e.g., “GDP” vs. “G.D.P.”), synonyms (e.g., “birth rate” vs. “fertility rate”), and annotated implicit references (Section 4.4.2) (e.g., “Companies with high revenue”), the pipeline attempts to resolve them with the following strategies:

- **Feature & Indicator:** We compute the semantic similarity between the indicator and the attributes from the data table using the *all-MiniLM-L6-v2* sentence BERT model and select the attribute with the highest similarity score.
- **Entity:** We apply semantic matching to entities without implicit references. Implicit references are resolved following the approach from existing work [160]. First, we use GPT-3.5 to convert them into SQL queries, which are then executed to retrieve the corresponding data. If SQL translation fails, we use the LLM’s comprehensive knowledge base to use an end-to-end question-answering process.
- **Date:** We first evaluate the expression using Python’s `eval()` function, setting *now* to the current year. For more complex date expressions, we employ techniques akin to those used for Entity resolution.

After accurately mapping the (Feature & Indicator, Entity, Date) tuple to the table’s corresponding attributes, we extract specific values. For instance, given the tuple (GDP, United States, 2010), we filter the table to exclusively present the GDP figures for the United States in 2010.

4.4.4 Statistical Check-worthiness

Apart from the two-stage pipeline, DataDive also supports identifying data-related and check-worthy sentences within the text the user is reading. This function is particularly useful for highlighting statistical statements in the UI, as detailed in Section 4.3.1. For assessing the check-worthiness of these statements, we use the ClaimBuster API [161], a web-based tool for live automated fact checking. Furthermore, to determine the relevance of the data, DataDive tailors a prompt for the GPT-4 model, following the approach of Liang et al. [162]. This prompt enables GPT-4 to provide a binary judgment on each sentence, ‘Y’ for relevance and ‘N’ otherwise. A statement is deemed highlight-worthy if it surpasses a threshold of 0.5 (empirically set) in check-worthiness and obtains a ‘Y’ in data relevance.

4.5 Pipeline Evaluation

To understand whether the pipeline of DataDive can produce meaningful questions for exploring contexts surrounding the statistical statements and to match the statistical statements to the relevant datasets capably, we conducted a technical evaluation of the pipeline. For the evaluation, we focused on statistical statements from online texts on global social issues. Therefore, we only had cases where statistical entities were countries and regions. We used the World Bank’s World Development Indicators [155] and Our World In Data’s Greenhouse Gas Emissions [153] and Energy Consumption [154] datasets.

4.5.1 Materials

In our study, we compiled 77 statistical statements and their contextual paragraphs from various online sources, including data journalism sections of The New York Times and The Economist, and social media platforms like Twitter, Reddit (r/changemyview, r/economics, r/news, r/dataisbeautiful), Quora, as well as blogs and non-governmental organization publications. We manually reviewed recent social media posts and news articles and searched for news articles and social media posts on the global population and climate change. From the articles and posts, we extracted statistical statements related to the World Bank’s World Development Indicators.

4.5.2 Participants

We recruited 18 external evaluators from university communities in South Korea, ensuring that they had no previous involvement or knowledge in the study. These evaluators were chosen based on their interest in social issues, experience searching for statistical information, and English proficiency to understand English texts on social issues. The evaluators were aged between 18 and 24 (Mean = 20.4), and were compensated with KRW 40,000 ($\approx$ USD 30).

4.5.3 Procedure

We randomly divided the statistical statements into three sets and assigned evaluators to each set. Each evaluator rated 25-26 statistical statements in total. Due to the dropouts, the number of evaluators per statement differed from five to seven.

We provided an evaluation system for the evaluators to read the statistical statement and its surrounding paragraph and examine the recommended questions and the set of indicators-year-country triplet matched to the statistical statement.

Then, the evaluator rated the quality of the pipeline results with 5-point Likert scale questions (1 = Strongly Disagree to 5 = Strongly Agree) on the (a) meaningfulness and the relevance of individual recommendations, (b) meaningfulness, diversity, and interestingness of the recommendation set, and (c) relevance of matched statistical entities, indicators, and time. The detailed questions are available on the appendix.

4.5.4 Results

The evaluation results showed that the pipeline could generate meaningful context recommendations for statistical statements and match them with relevant data sets.

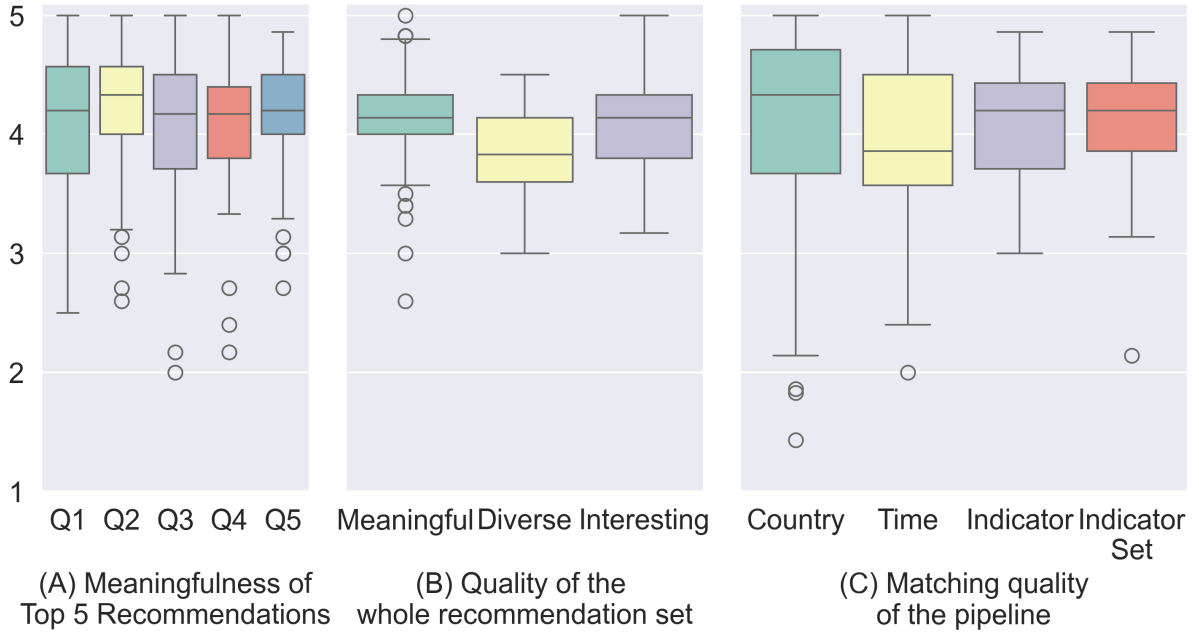


Figure 4.6: Boxplot of the average rating of recommendations and the matching quality from the pipeline. (A, B) The results show that evaluators considered the recommendations meaningful, diverse, and interesting. (C) While evaluators were generally favorable to the matching quality, matching for the country had the most errors.

Meaningfulness of Individual Recommendation Item

Figure 4.6A shows the evaluators' evaluation results of the top 5 recommendations. In general, participants considered that the recommendations generated by the pipeline would help readers gain meaningful insights into the statement's context. The average rating for meaningfulness for all recommendation questions was 4.12 ($SD = .55$). The rating was consistent for the top 5 questions, with average ratings between 4.08 and 4.20. We observed 13 instances where the average rating was lower than 3. Five cases involved recommending other countries with extreme values for statements on trends of an explicit set of countries. Three cases involved recommending years with extreme values of a statistical indicator for a single country for statements on comparing multiple countries. Other cases involved recommending unrelated statistical indicators or countries.

Also, the evaluators considered that the statistical entities mapped to each recommendation were relevant. Out of 385 recommendations with individual evaluations, only one case had a relevance score of less than 3.

Quality of Recommendation Set

Figure 4.6B shows the evaluators' evaluation results of the whole recommendation set. The average rating for the meaningfulness (4.12 / 5, $SD = .43$), diversity (3.86 / 5, $SD = .35$), and interestingness (4.07 / 5, $SD = .40$) of the recommendation set suggests that participants considered the recommendation set as useful.

Relevance of matched statistical entities/indicators/time

Figure 4.6C shows the participants’ evaluation results of the matching quality. Participants responded that the pipeline could get a set of data points relevant to the given statistical statement. The pipeline retrieved relevant countries and regions for 84.4% (65 out of 77) with an average rating of 4.02 / 5 ($SD = .90$), relevant time periods for 92.2% (71 out of 77) with an average rating of 3.92 / 5 ($SD = .69$), and relevant sets of statistical indicators for 100% (77 out of 77) of the cases with an average rating of 4.12 ($SD = .48$). A total of 80.5% (62 / 77) of the statements had relevant matching for all three components. The list of potentially relevant statistical indicators was rated as relevant for 98.7% (76 out of 77) of the cases.

We observed more errors for countries and regions, with two common types. The first type was when the statement did not directly refer to countries or regions, and the pipeline needed to infer the country from the context. The second type was when the statement involved extreme values and explicit references to countries, such as “Out of the large emitters the USA has the highest emissions per capita.” The pipeline retrieved the list of the top 5 countries in this case rather than the explicitly referred entity (USA).

4.6 User Evaluation

We conducted a user evaluation of DataDive to assess how DataDive supports the contextualization of statistical statements for users while they are reading texts and how it benefits the reading process. Specifically, we aimed to explore the following research questions:

RQ1 Does using DataDive support readers to explore more context around statistical statements from the text?

RQ2 Does using DataDive support the readers to gain more knowledge on the topic of the text?

RQ3 How do people cognitively/emotionally feel about using DataDive?

4.6.1 Participants

We recruited 21 participants with interests in and experiences with reading texts on social issues from universities in South Korea. For the screening, participants responded with their prior interests in global social issues on a 5-point Likert scale as well as open-ended responses on the specific topics of interest, and we filtered out participants without a high level of interest. Also, to ensure that the participants were able to understand English articles, we gave two short reading comprehension tasks in English and filtered participants with wrong answers. Participants (13 male, 8 female) were aged from 19 to 31 ($M = 22.9$) with diverse nationalities including South Korea ($N=12$), Kazakhstan ($N=3$), Bangladesh, Indonesia, Kyrgyzstan, Poland, Russia, and Spain (each with $N=1$). Each participant was compensated with KRW 30,000 ($\approx$ USD 22).

4.6.2 Conditions

We conducted a within-subjects experiment with two conditions:

- Baseline: Participants could use the web search freely, such as making queries to better contextualize content from the text they read. We selected the baseline to be close to the current behaviors.

- **DataDive:** In addition to using the web search, participants were able to use the full features of the system, including highlights on statistical claims, suggestions for contextualization, and data exploration.

4.6.3 Procedure

Participants read two texts for the DataDive condition and the Baseline condition, respectively. Participants were instructed to read the text thoroughly and encouraged to consider what information they wanted to learn more, especially related to their personal contexts. Before using the system, we briefly introduced it and let the participant use it for 10 minutes to get used to it.

Materials

We selected two articles from the New York Times on global issues, each with statistics. The articles were on carbon emissions⁷ (Carbon Emission) and decreasing fertility rate⁸ (Fertility Rate). To manage the reading time, the articles were reduced to around 500 words by removing various portions of the text.

Pre-Task

At the start of the session, participants filled out a survey with demographics and Likert-scale questions on their familiarity with online information search, online statistical information search, interest in carbon emissions, and interest in the population issue.

Task

The participants repeated the task twice, once with DataDive and once with only Baseline. The order of the texts and conditions was counterbalanced. In each task, participants were given 5 minutes to skim through the text and another 20 minutes to read it in detail. While reading the text in detail, they could search the web for external information to aid their understanding. Participants could also use DataDive depending on the experimental condition. During the session, we observed the participants' behaviors while reading the text. Thereafter, they responded to a post-task survey. We first asked open-ended questions about what they recalled from the text, the external information they explored, and the questions they had while reading. We further asked 7-point Likert-scale questions about their reading and information search experiences. Finally, we assessed the task load using NASA-TLX [163] and the system's usability using SUS [164] (for the DataDive condition).

Post-Task

After both tasks, we interviewed each participant. We asked participants about their workflow when conducting searches, what prompted them to search, and when they used the system. We also asked about their experience with the system and how it may have helped them contextualize the text. Participants were asked to reflect on how their reading experience with the system differed from how they might have read similar content in the past. Following this, we delved into their perceptions of the

⁷<https://www.nytimes.com/2016/04/06/upshot/promising-signs-that-economies-can-rise-as-carbon-emissions-decline.html>

⁸<https://www.nytimes.com/2023/02/09/upshot/china-population-decline.html>

Table 4.3: The average behavior counts to explore external information. The difference was significant for the Fertility Rate article and for the total.

	DataDive	Baseline	p -value	Stats
Fertility Rate	7.20	5.00	0.018	U = 21.5
Carbon Emission	7.36	6.00	0.195	U = 36.5
Total	7.29	5.48	0.017	W = 119

system, including what they liked or disliked, the issues they faced, and suggestions for improvement. We also asked about their willingness to use the system in the future and for what scenarios they would.

Data Analysis

To analyze user behavior, one of the authors counted the number of web searches and the number of instances of using DataDive in each session.

To analyze the interview responses, we first transcribed all interview recordings with ClovaNote⁹, and the two authors collaboratively extracted common themes from the interview transcripts.

To analyze participants’ open-ended responses to recalled facts after the task and to the recalled questions, we invited three external evaluators who were proficient in English and Korean. One evaluator analyzed both recalled facts and recalled questions, while the other two participated in either task. During the analysis process, two evaluators worked together to establish shared criteria. The evaluators created a list of recalled facts and a list of recalled questions per participant. Then, they annotated ones that were (a) about the data itself (e.g. specific values or trends) and (b) about the interpretation of data (e.g. data definitions, causal factors related to the trend, participants’ own insight). For recalled facts, evaluators additionally annotated ones that were from out of the text.

4.7 User Evaluation Results

In this section, we present how DataDive supported the contextualization of statistical statements and how it affected participants’ comprehension of data articles with statistical statements.

4.7.1 Assessment of RQ1

Our analysis of behavior logs and surveys suggested that DataDive led the participants to explore more external information, while the number of questions participants had did not show a significant difference across the conditions.

From the behavior logs, we observed that the participants attempted to search for more external information when using DataDive compared to the Baseline condition. (Wilcoxon Signed-Rank Test, $p = 0.017$, $W = 119$). The between-subject comparison showed that participants with DataDive searched more in the case of *Fertility Rate* article (Mann-Whitney U test, $p = 0.018$, $U = 21.5$), while the difference was not significant for the *Carbon Emission* article (Mann-Whitney U test, $p = 0.195$, $U = 36.5$) (Table 4.3).

⁹<https://clovanote.naver.com>

Table 4.4: The number of recalled questions per each article and condition. We did not observe statistically significant differences.

Article	Measure	DataDive	Baseline	p	Statistic
Carbon Emission	Questions	3.64	5.10	.830	U=51.5
	Data-related	2.09	2.10	.971	U=54.0
	Interpretation-related	2.45	2.00	.457	U=44.5
Fertility Rate	Questions	4.10	4.36	.190	U=36.5
	Data-related	1.70	1.73	.942	U=53.5
	Interpretation-related	1.30	1.36	.853	U=52.0
Total	Questions	3.86	4.71	.070	W=102.5
	Data-related	1.90	1.90	.903	W=79.5
	Interpretation-related	1.90	1.67	.750	W=54.0

Table 4.5: Participants' self-rating on exploring external information. There was no statistical significance between the DataDive condition and the baseline condition.

	Carbon Emission		Fertility Rate		Total	
	DataDive	Baseline	DataDive	Baseline	DataDive	Baseline
Motivated for searching	6.09	5.90	6.20	6.27	6.05	6.10
Success of finding information	4.64	5.30	4.70	5.45	4.71	5.38
Ease of finding information	4.91	4.90	4.90	5.18	4.90	5.05
Ease of understanding the information found	5.00	5.30	5.30	6.00	5.05	5.67
Trustworthiness of information found	5.55	5.70	5.70	4.91	5.62	5.29
Novelty of information found	5.09	5.00	4.90	4.45	4.81	4.71

From the survey, we compared the number of self-recalled questions for each condition (DataDive vs. Baseline) (Table 4.4). From the within-subject comparisons, we did not observe a statistically significant difference in the number of questions (Wilcoxon Signed-Rank Test, $p = 0.070$, $W = 102.5$ for total question count, $p = 0.903$, $W = 79.5$ for data-related question count, $p = 0.750$, $W = 54.0$ for interpretation-related question count), suggesting that DataDive supported the participants to explore more context per question. However, the self-perception of participants’ reading behaviors did not show significant differences between DataDive and the baseline condition (Table 4.5).

From interviews and observations, we found several instances of DataDive facilitating the exploration of more contexts around the statistical statements. We first introduce the common types of external information participants wanted and describe how DataDive’s underlines on statistical statements and recommendations of potential contexts supported exploring contexts around data.

During the tasks, participants had diverse motivations to search for external information in both the DataDive and baseline conditions. The most common case was searching for supporting evidence for statistical statements (10/21 participants). The simplest cases were checking whether the numbers or trends shown in the text aligned with the raw data they encountered. Participants also sought whether the trends from the data and their explanations were persistent and generalizable to other contexts, such as whether the decoupling between carbon emission and economic growth described in the *Carbon Emission* article was actually a prolonged trend or not (8/21 participants). Additionally, searches to gain more knowledge about the topic were common, such as the meaning of words (e.g., “Baby bonus” and “Speed premium” policy for boosting birth rates, “Paris Agreement” for reducing carbon emissions), and the definitions of the statistics mentioned in the text (8/21 participants).

In-text Underlines Nudging Exploration

Among the three ways to initiate exploration (underlined sentences, self-selected phrases, and entering into a search box), the underlined sentences were perceived to be more convenient and likely more reliable in ensuring that the system would provide high-quality results. 12 participants mentioned that DataDive’s underlined sentences on statistical statements nudged them to pay greater attention to the statement and be more interested in exploring the recommended questions for them. Out of 232 instances of using DataDive to explore statistical information, 141 instances (60.8%) started from clicking the underlined sentences and 61 instances from the free-form selections. P19 said, “As [underlined] highlights picked statistical statements, I could pay more attention to them, which I might have missed if I read [the article] without the system.”

Question Recommendations Encouraging Further Exploration

Participants commonly considered that the recommendation feature of DataDive encouraged them to explore the datasets. Out of 240 clicks on the recommended questions, 165 instances (68.8%) were clicks on the questions generated by the pipeline and 75 (31.2%) instances were clicks on the default question (“Does the data support this statement?”), showing that participants often clicked the recommended questions. Six participants valued the recommended questions as they provided novel aspects of the issue. Eight participants thought recommended questions on statistical indicators helped them expand their thoughts by considering novel aspects.

Among the recommended questions, personally relevant questions drew some participants’ interest to explore further. This was particularly prominent for questions mentioning the participants’ home and

Table 4.6: The number of recalled facts for each article and condition. The values are reported as DataDive vs. Baseline. We did not observe statistically significant differences between conditions, except for the number of recalled facts out of the text in the Carbon Emission article and in total.

Category	Type	Carbon Emission	Fertility Rate	Combined
Facts	All	5.00 vs. 7.20 ($p = .092$)	6.90 vs. 6.91 ($p = .972$)	5.95 vs. 7.05 ($p = .068$)
	Out-of-text	2.64 vs. 4.10 ($p = .008$)	2.50 vs. 3.45 ($p = .178$)	2.57 vs. 3.76 ($p = .011$)
Data	All	3.09 vs. 4.30 ($p = .144$)	4.80 vs. 4.36 ($p = .614$)	3.90 vs. 4.33 ($p = .282$)
	Out-of-text	2.50 vs. 2.00 ($p = .344$)	2.18 vs. 1.70 ($p = .347$)	1.86 vs. 2.33 ($p = .240$)
Interpretation	All	3.82 vs. 4.30 ($p = .825$)	4.60 vs. 4.09 ($p = .564$)	4.19 vs. 4.29 ($p = .791$)
	Out-of-text	2.36 vs. 2.20 ($p = .938$)	1.70 vs. 1.91 ($p = .589$)	2.05 vs. 2.05 ($p = 1.000$)

neighboring countries where they were interested to see what the situation reported in the text would be like in their own country (5/21 participants). The questions took the scope of their interpretation beyond the text to consider how it relates to themselves or other contexts they were aware of.

Nudging Features Potentially Leading to Passive Reading

However, at the same time, some participants (P3, P18) were concerned that the underlying and recommended questions would discourage them from thinking about their own perspectives and instead rely on the system’s guidance. P3 commented that “The list of questions made me read the text more passively. I could see the questions when I clicked the sentence, so I didn’t really think about the questions by myself. However, the provided questions were not really good ones touching the core of the issue, so I felt I was a bit trapped in some frame.” Participants also pointed out cases where the recommended questions were shallow and repetitive and where a greater variety would be appreciated (5/21 participants). P3 said that “The provided questions were too simple and specific. They were too focused on the facts themselves. I think the more important part is about the narrative with the numbers, such as cases where the fertility rate dropped even when policies for fertility rates were in action.”

4.7.2 Assessment of RQ2

Our analysis of the number of recalled facts and the self-evaluation on the knowledge gain did not show statistically significant results from using DataDive, but participants shared some beneficial cases from using DataDive.

From the post-task survey, we discovered that participants recalled more facts from the text and the information search in the Baseline condition than in DataDive condition (Wilcoxon Signed-Rank test, p

Table 4.7: Participants’ self-rating on their gain after the experiment. There was no statistical significance between DataDive and the baseline condition.

	Carbon Emission		Fertility Rate		Total	
	DataDive	Baseline	DataDive	Baseline	DataDive	Baseline
Knowledge gain	6.27	5.50	5.80	6.09	6.05	5.81
Critical attitude	5.73	6.30	5.30	6.09	5.52	6.19
Enjoyment	6.00	6.00	5.60	6.00	5.81	6.00
Personal relevance	5.09	4.80	4.40	5.36	4.76	5.10
Exposure to greater context	5.55	5.30	5.40	5.82	5.48	5.57

= 0.011, $W = 172.5$) (Table 4.6). However, the number of recalled facts on data or data interpretation did not show significant differences across conditions. This suggests that using DataDive led participants to recall relatively more data-related facts.

Participants’ self-evaluation on their perceived knowledge gain from the task did not show significant differences between the conditions (Table 4.7).

Still, we discovered notable cases where DataDive supported more discovery on the topic of the text from the interviews and observation. Participants generally felt that DataDive provided easy and flexible ways to explore data.

Increasing Awareness on Statistical Indicators

Some participants (P5, P12) valued that DataDive guided the use of data to observe social issues. P5 shared an example of how DataDive helped them decide which indicators to use, “When the text says ‘Countries A and B emit [carbon dioxide] a lot’, then I would have vague thoughts on seeing the carbon emissions of these countries and drag on the country names. DataDive suggested ‘CO2 emissions per capita’, and I realized that emission per capita would be a more relevant indicator in this case.” Similarly, P12 said, “Statistical statements often refer to the data abstractly, like ‘economic growth rate’. There are multiple different metrics for economic growth, and it’s hard to recall them even though I already know them. The system provides a list of relevant indicators, which helped me recall them.”

Also, with the variety of statistical indicators being available to the participants, participants had new discoveries by swapping and comparing data (7/21 participants). From the Fertility Rate article, P6 clicked on a statement on Sweden’s parental leave to discover a recommended question “What impact does the longer parental leave duration have on the labor force participation rate in Sweden?” From the matched datasets, P6 unexpectedly discovered the data on the ratio between female and male participation in the labor force, which was almost equal to parity. P6 commented that the data supported what they found about gender equality in Sweden from their previous web search, leading to more impressive knowledge.

Critical Examination of Text

The participants commonly used DataDive to verify the trends and values of data within the text (12/21 participants). They reflected that such verification behaviors were not common in their everyday reading, yet was helpful for reading the text more critically (9/21 participants).

Table 4.8: Participants’ task load for each condition with the NASA-TLX questionnaire. There was no statistical significance between DataDive and the baseline condition.

	Carbon Emission		Fertility Rate		Total	
	DataDive	Baseline	DataDive	Baseline	DataDive	Baseline
Mental Demand	2.73	2.80	2.70	3.09	2.81	2.95
Physical Demand	1.45	1.80	1.60	1.64	1.52	1.71
Temporal Demand	1.91	2.10	1.80	2.09	1.86	2.10
Performance	3.45	3.90	3.80	3.73	3.67	3.81
Effort	3.55	3.80	3.80	3.73	3.76	3.76
Frustration	2.09	2.10	2.40	2.36	2.33	2.24

Participants also mentioned that DataDive supported the identification of statistical fallacies by providing a larger context. P18 gave a specific example, “There are statistical fallacies, like the trend could be felt very different by how the data was normalized, or the data was relative or absolute. To be aware of such [fallacies] from the articles I have to read the text in detail and think critically. But with the system, I can discover such differences with a couple of clicks, which helped me critically read the statistics.” During the tutorial, P11 read a short piece of text claiming that automation did not cause the loss of jobs and by exploring the system, commented that “After I saw the actual data, I realized how the author was cherry-picking the data. The unemployment rate in the past was high due to the subprime mortgage crisis, and I thought the text was distorting the statistics. I wouldn’t have noticed it if I didn’t check the past data beyond what was written in the text.”

4.7.3 Assessment of RQ3

While the standardized metrics’ usability scores suggest that DataDive needs further improvement, participants generally valued the system for supporting focused browsing and interaction with data to contextualize the statement.

The average SUS score for the system was 74 (Min = 40, Max = 97.5, SD = 15.61), suggesting that participants had mixed opinions on the usability of the system. Task load between DataDive and the baseline condition did not show a significant difference, showing that using DataDive did not incur additional task load (Table 4.8).

Here, we list some commonly mentioned factors that affected participants’ perceptions of the system’s usability and usefulness.

Credibility of Provided Information

During the user study, DataDive showed data only from authoritative sources, with metadata on the data sources (Section 4.4.1). Participants typically considered the data provided by DataDive to be credible (9/21 participants).

Conciseness of Results

Where web search results tended to contain a variety of media and data formation from a diverse range of sources, the focused use case of DataDive was appreciated in removing the bloat of unneces-

sary content while still providing vital and meaningful information (4/21 participants). Furthermore, DataDive was appreciated for being more precise and efficient for statistical searches, as participants can gain access to and wrangle with the data immediately to fit their purpose, rather than having to sift through several web pages until they get to the relevant result or perhaps even fail to do so (4/21 participants).

Focused search

The structured nature of DataDive also allowed participants to be focused on diving into their exploration of the system as compared to web search, where the breadth of information may distract them and lead them to searches that deviate from their original intent (3/21 participants).

Easier Comparisons Between Data

Participants valued that DataDive provided easier interactions and options for freely comparing the data (11/21 participants). When searching for statistical data from the web, they commonly relied on image search or charts provided by the search engine, but felt limited as pre-generated charts show only a static set of indicators over pre-selected countries and time (5/21 participants).

4.8 Discussion

In this section, we discuss potential directions for future research in contextualizing statements with statistical data. First, we discuss additional application contexts that DataDive can support. Then, we discuss the design decisions of our LLM-powered pipeline to produce factually reliable yet diverse outputs. To generalize our findings, we discuss considerations for organically expanding the database of datasets to encompass diverse data sources. Lastly, we discuss the limitations of our study and future directions.

4.8.1 Incorporation of Contextual Information Beyond Structured Data

While DataDive focuses on supporting the discovery and exploration of relevant contexts based on statistical data in structured forms, contextual information in other forms (e.g., natural language text, visualizations, images) can introduce additional valuable information. In fact, a third of our user study participants mentioned that they would have appreciated contextual information beyond just the statistics in DataDive, such as governmental reports (P8) and news articles (P11). For example, information in Wikipedia¹⁰ or informational videos can provide general background knowledge on the topics [165]. Temporal information, such as important events related to statistical entities, can be augmented with data from news media, similar to the approach employed in Contextifier [63]. Furthermore, official reports from governments and global organizations as well as research articles can provide credible interpretations of data. Such additional information can be delivered to users through graphical overlays [166] and text annotations. For example, in Figure 4.1C, DataDive could provide annotations of important economic events in Sweden on the chart of GDP.

Including contextual information beyond structured data can also enable improvements in the recommendations generated by DataDive and its visualization interface by leveraging knowledge for generating recommendations of related entities or dates. For instance, in the example in Figure 4.1B, DataDive

¹⁰<https://en.wikipedia.org>

could suggest potential mediating variables between economic growth and fertility rate from external reports [167].

4.8.2 Considerations for Factually Reliable LLM-Powered System

In this work, we leveraged an LLM-based pipeline for two features: recommending potential contexts around statistical statements and mapping implicit data references and recommended contexts in statistical statements to existing statistical entities, indicators, and time periods. To design the two features, we utilized the generative capacity of LLMs in different ways.

For recommending potential contexts, we focused on diversity, using the LLM’s generation capability to create a range of potential contexts. On the other hand, for mapping the contexts to statistical entities and indicators, we prioritized relevance so that the pipeline would provide trustworthy information. Therefore, we avoided solely relying on LLMs for mapping data references and generated the names of entities using LLMs as a final resort. Instead, we matched references with semantic similarities first, and we used LLMs to generate a method for resolving entities, such as creating SQL statements to resolve the entities, to ensure more trustworthy responses. Since LLMs can hallucinate, we designed the system to produce a list of potential alternatives and allowed the users to make the final selection. To support users in making the decision, we also provided extra information such as teaser questions to explain how the recommended contexts relate to the text and the explanation of statistical indicators to describe how the contexts are related to the indicators.

However, for mapping implicit data references, we found that providing multiple candidates can cause more challenges. Errors related to countries and dates were easily noticed and recoverable, but participants had trouble recovering from incorrect statistical indicators as there were often many similar statistical indicators. With false positives, participants felt that it was difficult to determine whether the matched indicators were actually related to the statement, which, in the worst case, could have led to false interpretations. For false negatives, participants had difficulty deciding how to recover from the error as they were unsure whether the relevant indicators were present in the system or if their input statement was wrong. To address this, providing more accessible and concise information for the users to understand and verify DataDive’s output could better support them in handling errors. By communicating the errors and uncertain cases to users, for example by providing visual warnings with uncertain cases, users’ mental models of DataDive can also be improved.

4.8.3 Considerations for an Organically Growing Context Database

While we prepopulated the context database with datasets from Our World in Data [153, 154] and the World Bank [155], we envision that the support for contextualization by DataDive would apply to texts in other domains as well, such as local data [168]. To achieve this, users can potentially upload or provide links to their data based on their needs to DataDive, which would organically grow the database. We expect that this feature would attract a broader pool of users interested in domains not necessarily covered by our seed datasets.

However, we foresee several challenges in supporting the organic expansion of the system. Ensuring the credibility of the data [149, 150, 151] would be one challenge. Another issue would be ensuring the integrity and consistency of the data across different datasets, which is known to be one of the most effortful and time-consuming processes from data science literature [169, 170, 171]. Also, as DataDive encompasses more data, DataDive would need to be able to disambiguate a large set of indicators or

entities with similar names. One potential solution would be having moderators to manage the expanding datasets and build rules from data definitions and formats of existing centralized sources, such as Our World In Data ¹¹, Data Commons ¹², or FRED ¹³. With the established rules, moderators can manage the new datasets added to the system by validating against seed datasets and enforcing the matching format. We envision that such a collaborative ecosystem for organizing trustworthy and coherent datasets and supporting the collaborative analysis of data [172] will further enrich DataDive.

4.8.4 Limitations and Future Work

Generalizing Evaluation & User Study Findings. We evaluated the pipeline of DataDive with only text on global social issues and a database on global social issues. To verify the generalizability of the pipeline, future work can evaluate the pipeline with different sets of indicators and statistical entities from various contexts.

For the user evaluation, we only tested the system with news articles on social issues in a controlled setting. We expect that the motivations and behaviors of using DataDive would differ according to users' goals when reading text and the type of text. Also, in this study, we evaluated the system only with college students. We expect that the level of expertise and experience with data and the topic of the article would affect the value of DataDive. In future work, we expect that longitudinal studies in real-life settings with diverse participants, such as a diary study, could capture the usefulness of DataDive in diverse settings. Lastly, the recalled facts were not verified against the source article and the underlying data, so the measure does not distinguish accurate recall from misremembering; the counts in Table 4.6 should be read as an indicator of engagement with the material rather than of accurate learning.

Improving Runtime Efficiency. Due to the multiple calls to the LLM, the delay in responding to user input was often noticeable, which hampered the user experience (P4, P5, P6, P7, P9, P20). We believe that the runtime efficiency of the pipeline can be further improved by parallelizing calls to the LLM or leveraging more efficient LLM models with fine-tuning. Furthermore, preprocessing sentences while the user is reading through the text could reduce wait time.

4.9 Conclusion

The meaning of statistical statements differs by what context it was interpreted in. While readers may consider their own or relevant contexts when reading statistical statements to develop their own understanding, it is challenging to do so due to the barriers in searching for contextual information around them. We built DataDive, an interactive system to support the contextualization of statistical statements in texts by recommending diverse potential contexts around the statements and providing relevant data, powered by an LLM-based pipeline to parse statistical statements, generate potentially relevant contexts, and map statements to the datasets. Our evaluation results showed that the pipeline could produce meaningful recommendations and retrieve relevant datasets, and that the users of DataDive could easily explore contextualizations around statistical statements.

¹¹<https://www.ourworldindata.org>

¹²<https://www.datacommons.org/>

¹³<https://fred.stlouisfed.org/>

Chapter 5. FrameIt: Supporting Respectful Online Discussions through Moral Foundations

This chapter presents FrameIt. All uses of “we”, “our”, and “us” in this chapter refer to the collaborators who contributed to this work.

5.1 Motivation & Contributions

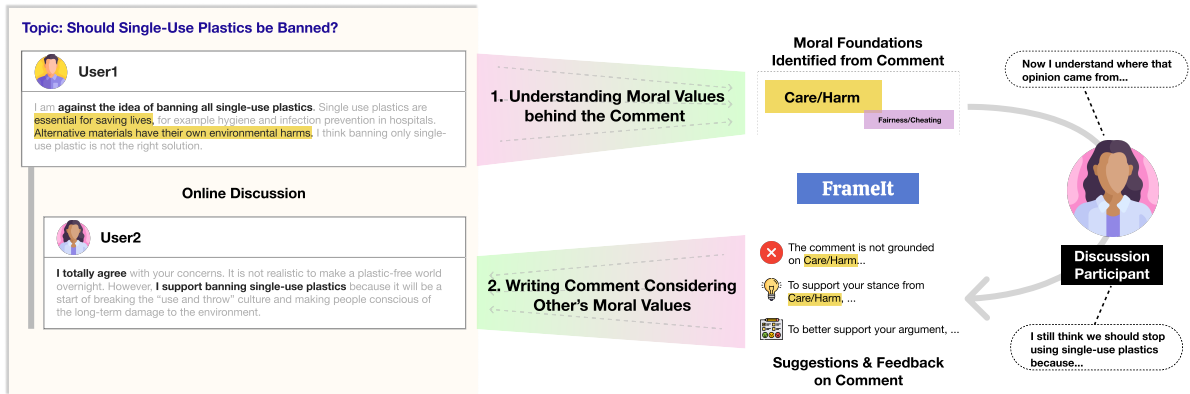


Figure 5.1: Conceptual diagram of FrameIt. FrameIt supports engaging in respectful online discussions by taking into account the moral foundations of the other discussion participants. From each comment, FrameIt identifies the moral foundations (e.g., Care/Harm) to help users better understand the perspectives of others. When writing comments, FrameIt provides feedback and suggestions that encourage users to consider the moral values of others.

Online discussions have the potential to be a space for constructive dialogue on social issues, in which participants exchange diverse perspectives and establish a shared understanding of the issues [112, 173]. Such dialogue is an essential element for informed public opinion and democratic decision-making. Realizing this potential requires people to acknowledge that people with different opinions may have valid reasons to disagree with their own positions [174] and to have a willingness to hear them out [175, 176], which we refer to “respectful engagement”. However, in reality, online conversations among participants with opposing views often fall short. People frequently end up talking past each other, failing to engage in a meaningful exchange of ideas, dismissing others’ views, and further reinforcing their own prior beliefs rather than considering different perspectives.

A key driver of this failure lies in how people interpret disagreements of opinions. People tend to experience their own views as direct, unmediated readings of reality. As a result, people tend to attribute disagreement to the other person’s ignorance or bias [177, 178, 179], rather than recognizing that opposing positions may stem from genuinely different underlying values [180]. This interpretive failure helps explain why interventions that simply increase exposure to opposing views often backfire, heightening incivility and polarization rather than reducing them [30, 181]. Exposure adds information about what others believe, but it leaves untouched how that belief is interpreted: encountering a view one already attributes to bad character only confirms the attribution. Fostering respectful engagement,

therefore, requires intervening not at the level of exposure but at the level of interpretation to shift how participants make sense of the disagreement they encounter.

Researchers have explored diverse interventions designed to foster respectful engagement. One thread of research focused on refining what views users encounter: instead of exposing people to opposing opinions indiscriminately, these systems provide algorithms and interactions to surface diverse yet agreeable opinions [60, 53, 51, 56, 57, 59, 58]. A second thread of research explored reshaping how opinions are expressed, nudging users to reflect on the quality and tone of their opinions [182, 183, 184, 185, 119] or directly rephrasing others’ messages to sound more receptive [173, 186]. Still, both threads leave the act of interpretation untouched, changing what users see or how they speak rather than how they make sense of a view once they encounter it. A third thread of research focused on scaffolding perspective-taking so that users can reason about why others hold a position. Existing research commonly provided social cues, such as demographic profiles, social media profiles, and social media feeds, to reason about the motivations behind different opinions [111, 75, 76]. However, identity-based cues risk oversimplifying individuals’ viewpoints or triggering partisan bias [180], and none of these approaches directly help users reinterpret the *source* of disagreements: the underlying values that give rise to it.

In this work, we investigate whether explicitly surfacing the differences in values underlying disagreements could facilitate respectful engagement. We hypothesize that translating surface-level opinion differences into differences in value systems can shift how users attribute disagreement, from a question of who reads the reality correctly to a question of which moral concerns each side weighs more heavily. We expect this reappraisal to foster recognition that an opposing position is grounded in a coherent set of value commitments, even where the reader does not share them, making the position more intelligible and more worth engaging with. We do not hypothesize that surfacing value systems will change participants’ opinions or make their arguments more persuasive; rather, the intended mechanism is a shift in how participants interpret and respond to disagreement and differing opinions.

We operationalize this approach with the framework of Moral Foundations Theory (MFT). The most widely used version of MFT describes five broad dimensions that shape people’s moral judgments through pairs of virtues and vices: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation [187, 188]. In this work, we selected MFT for three reasons. First, it offers an empirically grounded taxonomy for understanding the moral pluralism underlying variations in people’s opinions on social and political issues [187, 189, 190, 191, 192]. Second, research on “moral reframing”, a communication strategy of tailoring arguments to align with the moral foundations of the listeners, has shown that these foundations can facilitate cross-cutting communication on politicized issues [193, 194, 195, 196, 197]. Third, extensive computational research showed that moral foundations can be identified from texts [198, 199, 200] and used to generate morally grounded arguments [201], demonstrating the feasibility and validity of technical supports. We treat MFT as one productive instantiation of a broader idea of scaffolding discourse with a shared value framework, rather than as the only or the best set of values for shaping public reasoning.

However, it remains unclear how the concept of moral foundations can be leveraged to support laypeople’s everyday online discussions. To address this gap, we first conducted formative interviews with 13 frequent online discussion participants to understand their perceptions of moral foundations in online discussions and potential challenges of using them. Most interviewees imagined that MFT and moral reframing could be effective tools for respectful engagement, while raising practical concerns about: i) identifying others’ moral foundations from short comments, ii) constructing coherent arguments when reframing around unfamiliar perspectives, and iii) maintaining their sense of agency and ownership in

expressing their views and opinions.

Based on these insights, we developed FrameIt, a prototype interactive system that fosters respectful engagement with people with different opinions by explicitly surfacing other participants’ moral value systems (see Figure 5.1). While reading a discussion thread, FrameIt highlights the underlying moral foundations at both the comment and sentence levels, with explanations of how each foundation relates to the argument, surfacing the differences among moral foundations and their connections to the argumentation. To provide the annotations, FrameIt automatically analyzes each comment using a set of LLM prompts that generate both moral foundations annotations (as moral tags) and explanations. While replying, FrameIt supports the user in stating their points while being aware of differences in moral foundations. It offers three forms of support: a real-time indicator of moral alignment between the user’s reply and the target post’s foundations; LLM-generated discussion points that support the user’s stance while aligning with others’ foundations; and, upon request, LLM-generated feedback on drafts addressing moral alignment, logical validity, and tone.

We first evaluated the performance of LLM prompts in detecting moral foundations embedded in the comments using a subset of the Moral Foundation Reddit Corpus [202]. Our LLM prompts identified each moral foundation with precision ranging from 0.25 (Loyalty/Betrayal) to 0.53 (Care/Harm), recall ranging from 0.43 (Sanctity/Degradation) to 0.80 (Care/Harm), and F1 scores ranging from 0.34 (Loyalty/Betrayal) to 0.63 (Care/Harm). The F1 scores were comparable to the performance of the SOTA model. At the same time, these results indicate that the prompts tended to over-annotate moral foundations compared to human annotators, requiring critical interpretation of user responses.

To assess FrameIt’s effect on users’ perception towards different opinions and their discussion behaviors, we conducted two user studies: a within-subjects comparative study (N=22) examining FrameIt’s effect on users’ attitude towards different opinions, and a five-day-long deployment study (N=15) exploring FrameIt’s influence on the discussion atmosphere on a more holistic level. Our findings from the comparative study show that surfacing moral foundations helped participants better understand the value systems embedded in others’ comments, with the self-reported survey showing that participants felt they could better understand and sympathize with others. In the deployment study, participants described discussions with FrameIt as more respectful compared to their experiences on typical online discussion platforms. They specifically reported that moral foundation labels facilitated perspective-taking and reflection on their own arguments. At the same time, we observed an important tension: some participants perceived that FrameIt blurred productive contestation, as awareness of fundamental value differences sometimes led to premature “agreeing to disagree” rather than deeper engagement. These findings surface a meaningful design tension between deliberative and agonistic ideals for online discourse arising from the design of FrameIt

Summing up, this chapter makes the following contributions:

- FrameIt, an interactive prototype system for surfacing moral foundations to scaffold respectful engagement with different opinions,
- Empirical findings on how surfacing moral foundations can shape perceptions of different opinions and discussion behaviors, revealing both benefits for understanding and receptiveness towards opposing views and a potential tension between fostering respectful engagement and maintaining productive contestation.

5.2 Formative Study

To identify practical and perceptual barriers to using moral foundations in online discussions, we conducted formative interviews with frequent users of online discussion platforms. The interviews focused on two questions: (a) how people perceive Moral Foundations Theory and what challenges they anticipate, and (b) the appropriate level of technical support and reliance for respectful engagement. Although MFT is rooted in psychology, we deliberately recruited frequent discussion participants rather than expert psychologists: our goal was a prototype grounded in users’ authentic needs and expectations, not a flawless psychological instrument.

We recruited 13 participants who engaged in online discussions on social issues at least three times per week for a 1-hour semi-structured interview. Recruitment channels included a local university board, X (formerly Twitter), Reddit, Upwork, and Prolific. Participants’ ages ranged from 19 to 61 and represented diverse nationalities, including the US, Kenya, and India. Their primary discussion platforms were Reddit, X, and Facebook. In the interviews, we first asked about participants’ motivation for participating in online discussions and the main challenges that they encountered in pursuing constructive conversations. We then introduced the concept of moral foundations and moral reframing, asking them to imagine how they might apply it and what challenges they anticipated with this approach. Finally, we asked about participants’ perceptions of AI support and reliance in online discussion spaces. We analyzed the interviews using affinity diagramming and identified three common themes that led to our design goals.

5.2.1 Design Goals

In the interviews, most participants described online discussions as a space to share opinions, learn something new, and understand others’ viewpoints, rather than to persuade. While few were familiar with Moral Foundations Theory or moral reframing, they valued moral reframing as a way to foster respectful conversation by understanding and caring for others’ value systems. From these insights, we derive three design goals for applying MFT to facilitate respectful online discussions.

DG1. Support the identification of the moral foundations of other discussion participants.

Participants commonly expressed concerns about their ability to understand others’ moral values from limited information about other participants in the discussion. F10, F12, and F13 specifically pointed out the difficulty of grasping other people’s moral foundations from limited information available in online discussion spaces, which is often limited to comments and profiles (such as on Reddit). They often noted that the information displayed is not enough to understand other discussion participants and where they are coming from. To address this, we propose that the system should enable users to easily identify moral foundations from the comments. Because identifying moral foundations involves subjective interpretation and remains challenging for automated methods [110], the system should provide reasoning at the comment and sentence levels to explain its results and allow users to correct its mistakes.

DG2: Support aligning one’s arguments with the moral foundations of other discussion participants

Another challenge that was mentioned was linking the specific moral foundations to the arguments. F13 mentioned the challenge of connecting the moral foundations with the specific discussion topic, “Connecting moral foundations with arguments might be hard. Even if I know that there are some commonalities that I have (with others), they might be too far from the discussion topic.”

Most participants (F1, F2, F6, F8, F9, F10, F11, F12, F13) considered logically concrete arguments with factual support to be essential for a constructive discussion. At the same time, we expect that framing arguments around others’ moral foundations could complicate the construction of logically coherent arguments from differing perspectives. Therefore, we propose active, specific support to align their arguments with others’ moral foundations while maintaining logical coherence, fostering a constructive and respectful discussion environment.

DG3: Preserve users’ sense of agency in shaping arguments Participants valued online discussion as a space to express their own opinions while listening to others’ views, and emphasized the importance of maintaining agency over their contributions. F11 worried that moral reframing could feel like “Opening a bit too much to the other side”, and requiring agreement beyond their comfort. This concern aligns with prior findings that moral integrity can limit people’s willingness to apply moral reframing [203]. Participants also expressed that the appropriate level of system support should remain to assist the argumentation process, such as providing background knowledge (F8), counterarguments (F7, F11, F13), or advice on language choice (F1, F2, F10, F11), rather than the AI directly shaping their arguments by explicit interventions. Accordingly, we propose that system support should remain at the level of suggestions and feedback, preserving users’ agency in shaping their own arguments.

5.3 FrameIt

FrameIt is intended for individuals who participate in discussions on morally charged topics not to reach common ground or persuade others, but to exchange views and understand new perspectives. Rather than steering participants toward agreement, FrameIt fosters a respectful environment by surfacing the moral frames underlying each position. We illustrate the overall interactions with FrameIt through a hypothetical use-case scenario, followed by implementation details for moral foundation detection and for suggestion and feedback generation.

5.3.1 Use Case Scenario

Amy Belkins is a 28-year-old married freelancer with two kids. She loves discussing topics online, especially social issues, to exchange various perspectives and learn more about them. One of the topics she is most passionate about is the ‘Single-Use Plastic Ban’. She believes that single-use plastics should be banned due to the environmental harm they pose.

Reading Comments with Moral Foundations of Others (DG1)

Amy discovers a thread on FrameIt that argues against the single-use plastic ban. She opens the thread and reads the comments from other FrameIt users who are arguing against her opinion. While reading these comments, she notices tags of moral foundations on each comment header, as well as on a sentence-level (Figure 5.2A). From reading the comments, she finds that most comments that argue against the ban have the “Care/Harm” tag. Intrigued by the tags, she hovers on the “Care/Harm” tags on each comment to read how the moral foundation of “Care/Harm” relates to the comment (Figure 5.2B). After reading them, she learns that most of the comments focus on the necessity of plastics, such as for hygiene, and the potential economic damage (DG1). She understands that people who oppose the ban on single-use plastic are concerned about the harm to people. At the same time, she thinks the existing

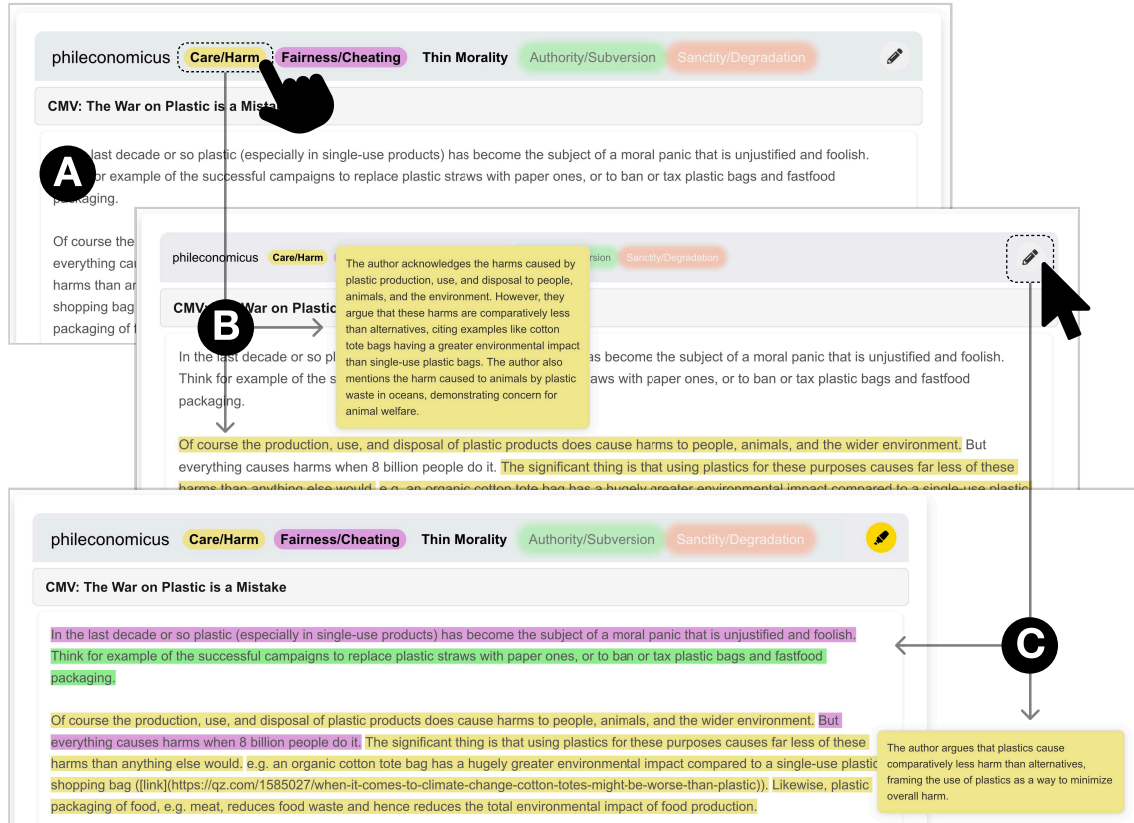


Figure 5.2: FrameIt presenting the comments in the discussion thread. (A) FrameIt presents a list of moral foundations relevant to the comment as tags. Foundations with high confidence have a solid outline, while those with low confidence have a diffuse outline. (B) When hovering on one moral foundation, FrameIt explains how it is embedded in the comment as a tooltip. FrameIt also highlights sentences related to that foundation. (C) When the highlight button is clicked, FrameIt highlights each sentence with its moral foundation. When hovering on each sentence, FrameIt explains how each sentence is related to the moral foundation.

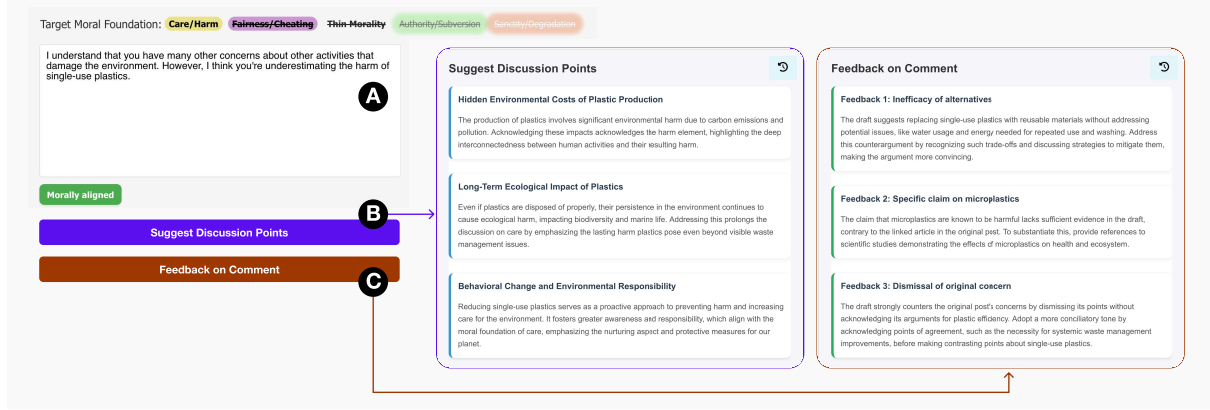


Figure 5.4: User interaction of FrameIt for using suggestions on discussion points and feedback. (A) The user can write their comment in the textbox. (B) When the user clicks Suggest Discussion Points, FrameIt generates a list of potential discussion points based on the original post’s content, the current comment content, and the target moral foundations. (C) When the user clicks the Generate Feedback button, FrameIt generates a list of feedback on logical flaws, misalignment of moral foundations, and the tone of the comment.

plastic ban from the “Care/Harm” perspective, including rephrasing environmental damage as long-term harm to humans and suggesting a compassionate transition (Figure 5.4B). Amy thinks the ideas would be effective for moving the discussion forward while respecting the OP’s opinion (DG3). She iterates her initial comment into a fully written text.

While writing, FrameIt consistently diagnoses the moral foundations embedded in the draft in real-time and visually indicates whether her draft is aligned with the “Care/Harm” foundation with the *Moral Alignment Indicator* (Figure 5.3C). When she tries to conclude her argument by emphasizing the importance of preserving the environment’s purity, the indicator shows “Morally unaligned”. When hovering over it, Amy learns that the argument about purity concerns the “Sanctity/Degradation” foundation and iterates the tone of the argument (DG2, DG3).

After finishing the initial draft, she wants to make sure her comment is well-supported and well-received by other discussion participants (DG2). Amy clicks on the “*Feedback on Comment*” button. FrameIt provides a list of ideas to improve the comment, including potential counterarguments and ideas for better aligning with the target moral foundation (Figure 5.4C). With the feedback, Amy revises the comment several times, gains confidence in the comment as she believes it will push the discussion forward, and posts it for others to see.

5.3.2 System Implementation

FrameIt employs an LLM-powered pipeline with two components: 1) detecting the moral foundations in text and 2) generating argumentation support and feedback grounded in these moral foundations. Although there are different versions of the MFT [204, 187, 205, 206, 188], we have decided to design our system with the original version described by Haidt and Joseph [187, 188] as most datasets and benchmarks have been designed around this version of the MFT [202, 198]. This choice lets us reuse existing resources for both implementation and evaluation.

Detecting Moral Foundation from Comments in FrameIt

To identify moral foundations in comments, we designed a set of LLM prompts. We chose to use LLMs instead of existing moral foundation classifiers [198] for their potential to generate explainable results. Specifically, our prompts were structured to provide sentence-level classification and natural-language rationales for each detected moral foundation, enabling users to critically assess and understand the system’s results (DG1). Each classification is done in the scope of the comment. This design choice was made to make the classification as objective as possible, since the holistic input from the discussion might introduce bias, especially given that the input tokens might be too large to analyze holistically.

As the detection of moral foundations is inherently subjective, we aimed to incorporate diverse perspectives in determining them. To that end, we employed two state-of-the-art LLMs: FrameIt first uses the `gpt-5-2025-08-07` model to identify an initial set of relevant moral foundations for a given comment. Then, a `claude-sonnet-4-20250514` model identifies another set of relevant foundations from an alternative perspective, generating a rationale for each moral foundation, and produces sentence-level annotations. If a moral foundation is detected by both models, it is annotated as *likely*. If it is detected by only one model, it is annotated as *moderate*. The *likely* foundations are presented with a solid outline in the interface, while the *moderate* ones are presented with a diffuse outline (Figure 5.2).

During our iterations, we found that GPT models tended to favor precision, assigning moral foundations only under clear relevance, while Claude models favored recall, assigning them more broadly. By combining these models, we aimed to produce a comprehensive set of relevant moral foundations while still distinguishing the most strongly supported. With this, we believe that FrameIt provides a good basis for constructive and respectful discussion through showing the foundations in a non-binary representation.

Supporting Argumentation with Moral Foundation

To support the writing process of arguments grounded in moral foundation with the goal of fostering a constructive and respectful environment, FrameIt provides three main features: (1) real-time analysis of moral alignment, (2) discussion point suggestions, (3) and feedback generation.

For the *Moral Alignment Indicator*, FrameIt analyzes the moral foundations of the user’s comment in real time and compares them with those of the original comment. FrameIt uses the same prompt for moral foundation detection with `gpt-5-2025-08-07`, producing a verdict on the alignment.

The *Suggest Discussion Points* feature is powered by `gpt-5-2025-08-07`, with a few-shot prompt designed to a) identify the main claim and the stance on the issue from the user’s current draft and b) generate discussion points that support the user’s main claim while aligning with the target moral foundations. This design helps users consider the moral perspectives of others (DG2) without altering their main claim (DG3).

To implement the *Generate Feedback* feature, FrameIt uses `gpt-5-2025-08-07` with a prompt to generate feedback on improving the comment draft. The feedback addresses five types of issues: (i) potential counterarguments, (ii) logical fallacy, (iii) lacking evidence, (iv) moral misalignment, and (v) tone of the comment. The first three (i.e., potential counterarguments, logical fallacy, lacking evidence) are adapted from prior work on LLM-supported argumentation tools [91], while the latter two were added to support our goal of fostering constructive arguments through moral reframing.

Technical Stack

FrameIt was built as a standalone prototype system using React in the frontend, and FastAPI as backend.

5.4 Technical Evaluation

We conducted a technical evaluation using an existing dataset to assess whether our pipeline can reliably identify moral foundations in online comments. As determining the relevant moral foundation is an inherently subjective task and lacks a definite ground truth, we do not expect perfect agreement with the ground truth data. Instead, this evaluation provides an estimate of the level of reliability users’ might experience when interacting with FrameIt.

5.4.1 Methods

As a ground truth dataset, we used the Moral Foundations Reddit Corpus [202], which contains 16,123 English Reddit comments. Each comment was annotated by at least three human coders for moral foundations. Annotations fall into three categories: (i) a list of all related moral foundations, (ii) *thin morality*, a moral claim without a clear moral foundation, or (iii) *non-moral*, i.e., no moral claim was found. Each annotation was also assigned a three-level confidence label: *confident*, *somewhat confident*, and *not confident*.

Because the dataset does not provide a single consensus annotation per comment, we first aggregated the annotations. For each moral foundation, as well as *thin morality* and *Non-moral*, we computed an aggregated likelihood score by weighting annotations according to the annotator’s confidence using a weighted average. We assigned a weight of 5 to confident annotations, 3 to somewhat confident, and 1 to not confident, respectively. The score for each label was computed as:

$$\text{Score} = \frac{(5 \times \text{Confident} + 3 \times \text{Somewhat Confident} + \text{Not Confident})}{(\text{Number of Annotators})}$$

For each moral foundation, we then decided an intensity label based on the score: *likely* if the score exceeded 1.67, *moderate* if the score was below 1.67, and *unlikely* if no annotator agreed on it. The threshold of 1.67 was empirically selected from the score distribution, corresponding to the case where at least two annotators, including one *confident* annotator, agreed out of three.

Table 5.1: The number of occurrences per moral foundation and the likelihood in the ground truth.

Likelihoods	Extremely Likely	Moderate	Extremely Unlikely
Care/Harm	168	235	1097
Fairness/Cheating	170	261	1069
Loyalty/Betrayal	31	130	1339
Authority/Subversion	60	229	1211
Sanctity/Degradation	18	127	1355
Thin Morality	177	505	818
Non-Moral	837	324	339

Table 5.2: The precision, recall, and F1 score per each moral foundation. Our evaluation results show that the our LLM prompts showed performance comparable to the SOTA model.

Frame	Model	Precision	Recall	F1	F1 Binary score of SOTA Model [198]
Care/Harm	GPT-4o + Claude-3.5-Sonnet	0.53	0.73	0.62	Care: 0.50 / Harm: 0.56
	GPT-5 + Claude-4-Sonnet	0.53	0.80	0.63	
Fairness/Cheating	GPT-4o + Claude-3.5-Sonnet	0.38	0.85	0.52	Fairness: 0.57 / Cheating: 0.61
	GPT-5 + Claude-4-Sonnet	0.46	0.74	0.57	
Authority/Subversion	GPT-4o + Claude-3.5-Sonnet	0.29	0.83	0.43	Authority: 0.39 / Subversion: 0.37
	GPT-5 + Claude-4-Sonnet	0.37	0.67	0.48	
Loyalty/Betrayal	GPT-4o + Claude-3.5-Sonnet	0.22	0.70	0.33	Loyalty: 0.64 / Betrayal: 0.40
	GPT-5 + Claude-4-Sonnet	0.25	0.53	0.34	
Sanctity/Degradation	GPT-4o + Claude-3.5-Sonnet	0.27	0.43	0.33	Sanctity: 0.49 / Degradation: 0.25
	GPT-5 + Claude-4-Sonnet	0.34	0.43	0.38	
Thin Morality	GPT-4o + Claude-3.5-Sonnet	0.47	0.44	0.45	N/A
	GPT-5 + Claude-4-Sonnet	0.51	0.30	0.37	

From the dataset, we randomly sampled 1,500 comments (i.e., approximately 10%), to ensure that our evaluation covers a representative number of samples with reasonable API costs. Table 5.1 shows the ground truth distribution of the moral foundations in the sampled comments.

We measured the precision and recall for each moral foundation, treating both *likely* and *moderate* as positive cases. Precision was defined as the proportion of cases that the pipeline labeled as positive that were also labeled as positive in the ground truth. Recall was defined as the proportion of positive cases in the ground truth that were also identified as positive by the pipeline.

5.4.2 Results

We evaluated the performance of the pipeline using two criteria: (i) precision and recall in detecting non-moral comments and (ii) the likelihood scores for each moral foundation.

Table 5.2 reports the precision, recall, and F1 score for each moral foundation across the two generations of flagship LLMs available at the time of our studies. With `gpt-5-2025-08-07`, the precision ranged from 0.25 (Loyalty/Betrayal) to 0.53 (Care/Harm), and the recall ranged from 0.43 (Sanctity/Degradation) to 0.80 (Care/Harm). Our prompts achieved F1 scores on par with MoralBERT [198], the state-of-the-art model for detecting moral foundations, even though the evaluation setups differ: our pipeline classifies five moral foundations, whereas MoralBERT tackles a more challenging 10-class task by separating each foundation into its virtues and vices (e.g., Care vs. Harm). Our empirical results suggest that our prompts tend to assign moral foundations to the comments more readily than human annotators. We consider this an acceptable bias, reflecting a tradeoff between precision (i.e, risk of omitting some moral foundations) and recall (i.e., risk of including less relevant moral foundations). In FrameIt, the pipeline deliberately prioritizes recall, since users can critically assess the system’s analysis and disregard irrelevant moral foundation labels. This design choice aligns with our design goal of FrameIt being supportive but not intrusive (DG3).

We also observed variation in precision across moral foundations: our LLM prompts more reliably detected Care/Harm and Fairness/Cheating compared to other foundations. We attribute this to the underlying language model having more training data for Care/Harm and Fairness/Cheating [207, 208].

Table 5.3: Demographic profiles of the user study participants and their prior opinions on the two issues used in the experiment. The prior opinions were measured in a 7-point Likert scale (1: Strongly Disagree, 7: Strongly Agree).

No.	Age	Gender	Nationality	Single-use Plastic Ban		Cancel Culture	
				Prior	Condition	Prior	Condition
CP1	28	Female	South Korea	5	FrameIt	6	Baseline
CP2	24	Female	South Korea	6	FrameIt	3	Baseline
CP3	28	Female	Peru	5	Baseline	5	FrameIt
CP4	22	Female	Philippines	6	Baseline	4	FrameIt
CP5	21	Male	Azerbaijan	5	FrameIt	4	Baseline
CP6	23	Male	Austria	3	Baseline	1	FrameIt
CP7	24	Male	Austria	6	Baseline	1	FrameIt
CP8	24	Female	South Korea	6	FrameIt	4	Baseline
CP9	26	Male	South Korea	6	Baseline	2	FrameIt
CP10	24	Male	Kazakhstan	5	Baseline	4	FrameIt
CP11	21	Male	Myanmar	6	FrameIt	3	Baseline
CP12	26	Male	Bosnia and Herzegovina	6	FrameIt	2	Baseline
CP13	22	Female	South Korea	4	Baseline	2	FrameIt
CP14	21	Female	South Korea	6	FrameIt	4	Baseline
CP15	28	Female	South Korea	6	Baseline	4	FrameIt
CP16	26	Female	South Korea	5	Baseline	2	FrameIt
CP17	24	Male	South Korea	5	FrameIt	2	Baseline
CP18	21	Female	Kazakhstan	6	FrameIt	4	Baseline
CP19	29	Male	South Korea	2	Baseline	2	FrameIt
CP20	22	Male	Kyrgyzstan	5	FrameIt	4	Baseline
CP21	21	Male	Tajikistan	4	FrameIt	2	Baseline
CP22	19	Male	South Korea	5	Baseline	2	FrameIt

5.5 User Study 1: Comparative Lab Study

To examine whether engaging in online discussions with FrameIt fosters more respectful opinion expression, we conducted a synchronous, comparative lab study. The study focused on assessing the effects of FrameIt on both reading and writing comments, compared with conventional online discussions.

5.5.1 Participants

We recruited 22 participants from the online community of several universities in South Korea and Austria. Participants (12 male, 10 female) were aged 21 to 29 ($M = 24$) and represented diverse nationalities. Each received KRW 40,000 (approximately USD 30) as compensation for completing the study. Table 5.3 summarizes participant profiles, including age, gender, nationality, and prior opinions on the issues used in the experiment. This study received approval from a university’s Institutional Review Board (IRB).

5.5.2 Conditions

We conducted a within-subjects study with two conditions: FrameIt and a baseline. In the FrameIt condition, participants had access to all features of FrameIt powered by the most up-to-date models available at the time of the study, `gpt-4o-2024-08-06` and `claude-3-5-sonnet-20240620`. In the baseline condition, only a simple input box was provided to simulate a conventional online discussion setting. The within-subjects study design allowed us to account for people’s different discussion habits while also collecting participants’ qualitative comparisons between FrameIt and conventional online discussion methods.

For the study, we selected two topics that a) had a variety of opinions based on a diverse set of moral foundations, b) were well-known to the public to ensure the participants had initial thoughts on the issue, and c) were not specific to a single country. We selected two topics from ProCon.org¹: *Single-Use Plastic Ban* and the effect of *Cancel Culture*. We counterbalanced the order of the conditions and the discussion topics, while ensuring equal distribution of prior opinions across conditions.

To simulate a realistic discussion setting for each topic, we searched Reddit’s r/changemyview for relevant discussion threads. We used the following discussion thread for each topic:

- Single-Use Plastic Ban: The War on Plastic is a Mistake²
- Cancel Culture: There is no remedy for “Cancel Culture” that doesn’t involve the destruction of freedom of expression³

To ensure coverage of diverse moral foundations while balancing cognitive load across the topics, we selected a subset of comments from each thread that represented a range of perspectives on the issue. For the discussion on *Single-Use Plastic Ban*, we included the original post and seven (out of 98) comments, and for *Cancel Culture*, the original post and five comments (out of 444). The difference in the number of comments is attributed to differences in comment length and in the distribution of moral foundation tags, ensuring a similar spread of total comment length and diversity of foundations represented. To observe the use of FrameIt for diverse moral foundations and contrasting viewpoints, we also selected one comment per topic that explicitly opposed the stance of the original post. In the study, participants were asked to reply to both the original post and the selected opposing comment.

5.5.3 Procedure

Participants first completed a short pre-task survey about their prior opinion on the issues and their perceptions of people with incongruent views.

They then received a short introduction to the Moral Foundations Theory and moral reframing. We introduced the moral foundations as “a set of foundational values that trigger people’s moral thoughts, universally observed across different cultures and populations” [187], and described each foundation in detail based on the definitions by Haidt [187]. Moral reframing was introduced as “a persuasive technique that frames the argument based on the moral foundation that others care about”, illustrated with an example from prior work [203].

For each discussion topic, participants read the discussion thread and posted one reply to the original post and one reply to a pre-selected comment, simulating typical online discussions. Each topic

¹<https://www.britannica.com/procon>

²https://www.reddit.com/r/changemyview/comments/17mtjoz/cmv_the_war_on_plastic_is_a_mistake/

³https://www.reddit.com/r/changemyview/comments/1cx9bix/cmv_there_is_no_remedy_for_cancel_culture_that

was assigned 30 minutes. Participants were free to use external resources, including web searches and LLMs. During these sessions, we observed them engaging in reading and writing comments.

After discussing each topic, participants completed a short post-task survey about their perceptions of incongruent views and their discussion experience.

After completing all tasks, we conducted a semi-structured interview focusing on their experience with FrameIt, including how using FrameIt to read and write opinions differed from using the baseline, how they used specific features, and potential improvements.

5.5.4 User Study Measures

To evaluate changes in participants’ attitude towards the opinions of others, we adopted an existing set of questions on conversational receptiveness [175, 209]. From the original 18 questions, we selected 10 measuring intellectual curiosity and derogation of opponents, excluding those on ‘negative emotions towards incongruent views’ because they overlapped with other questions, and ‘feeling issues as taboo’, which were more about the perception of the issue itself.

For qualitative insights, we transcribed interview logs using ClovaNote⁴ and collaboratively developed themes of system usage through author discussion. To complement these findings, we also asked 7-point Likert-scale questions on the users’ discussion experience, focusing on their attitude towards others while reading and writing opinions.

5.6 User Study 1 Results

In this section, we report our findings of how using FrameIt affected participants’ attitudes and behaviors toward commenters holding different opinions. We also report our observations on how participants used FrameIt.

From a within-subjects comparison of self-reported responses of discussion experience, we found that the participants could better understand the opinions of others (Wilcoxon Signed-Rank test, $p = 0.0016$, $W = 8.0$) and sympathize with them (Wilcoxon Signed-Rank test, $p = 0.0089$, $W = 22.5$) when using FrameIt compared to the baseline interface (Table 5.4). We also conducted a between-subjects comparison of the scores across conditions for each topic. On the *Cancel Culture* issue, participants using FrameIt reported better understanding of others’ opinions than those in the baseline condition (Mann-Whitney U test, $p = 0.038$, $U = 91.5$), whereas no statistically significant difference was observed for the *Single-Use Plastic Ban* issue.

Qualitative evidence from the interviews further revealed that FrameIt helped participants understand others’ opinions and sympathize with them. Participants described the experience with FrameIt as reading the comment from the commenter’s perspective rather than their own, which gave them a better sense of understanding others and sympathizing with them. As CP15 explained, “With the analysis [of FrameIt], I thought, ‘This person thinks this value is very important, so that’s why they are making this point,’ giving me a sense of understanding others better. Without the system, I just superficially understood the arguments.” In contrast, when using the baseline system, participants reported difficulties identifying the moral foundations underlying other’s comments, which limited their ability to understand others’ values and reasoning.

⁴<https://clovanote.naver.com>

Table 5.4: Pairwise comparison of self-evaluated discussion experience between FrameIt and baseline conditions. Questions with statistical significance were bolded.

Discussion Experience in General	Baseline	FrameIt	p / W
The arguments from the comments were worth reading.	5.77	6.09	p = 0.14 W = 10.5
From reading the discussion, I deeply understood the opinions of others including their underlying motivations and values.	5.00	6.09	p = 0.0016* W = 8.0
From reading the discussion, I could sympathize with the opinions of others.	4.45	5.45	p = 0.0089* W = 22.5
I felt full ownership over the opinions I shared in my comments.	5.95	6.09	p = 0.58 W = 27.0
I wrote comments that were logically concrete and well-supported with evidence.	4.54	4.86	p = 0.36 W = 38.0
I wrote comments while respecting the views of the person I replied to.	5.45	5.82	p = 0.21 W = 23.5
After reading my comment, I believe the person I replied to would sympathize with my view on the issue.	4.55	5.14	p = 0.075 W = 29.5
I felt like my comment would be able to convince other people with opposing views on this issue.	4.86	4.95	p = 0.85 W = 64.5

Table 5.5: Comparison of self-evaluated discussion experience between FrameIt and baseline conditions on Single-use Plastic Ban issue. We did not observe any statistically significant difference.

Discussion Experience on Single-Use Plastic Ban	Baseline	FrameIt	p / U
The arguments from the comments were worth reading.	5.72	6.18	p = 0.33 U = 74.0
From reading the discussion, I deeply understood the opinions of others including their underlying motivations and values.	4.91	6.0	p = 0.054 U = 89.0
From reading the discussion, I could sympathize with the opinions of others.	4.90	5.64	p = 0.21 U = 79.0
I felt full ownership over the opinions I shared in my comments.	5.72	6.09	p = 0.56 U = 69.5
I wrote comments that were logically concrete and well-supported with evidence.	4.27	4.64	p = 0.31 U = 76.0
I wrote comments while respecting the views of the person I replied to.	5.64	6.18	p = 0.19 U = 80.0
After reading my comment, I believe the person I replied to would sympathize with my view on the issue.	4.45	5.27	p = 0.085 U = 86.0
I felt like my comment would be able to convince other people with opposing views on this issue.	4.54	4.82	p = 0.76 U = 65.5

Table 5.6: Comparison of self-evaluated discussion experience between FrameIt and baseline conditions on Cancel Culture. Questions with statistical significance were bolded.

Discussion Experience on Cancel Culture	Baseline	FrameIt	p / U
The arguments from the comments were worth reading.	5.82	6.0	p = 0.65 U = 67.5
From reading the discussion, I deeply understood the opinions of others including their underlying motivations and values.	5.09	6.18	p = 0.038* U = 91.5
From reading the discussion, I could sympathize with the opinions of others.	4.00	5.27	p = 0.07 U = 88.0
I felt full ownership over the opinions I shared in my comments.	6.18	6.09	p = 0.97 U = 61.5
I wrote comments that were logically concrete and well-supported with evidence.	4.82	5.09	p = 0.68 U = 61.5
I wrote comments while respecting the views of the person I replied to.	5.27	5.45	p = 0.59 U = 69.0
After reading my comment, I believe the person I replied to would sympathize with my view on the issue.	4.64	5.00	p = 0.22 U = 78.5
I felt like my comment would be able to convince other people with opposing views on this issue.	5.18	5.09	p = 1.0 U = 60.5

We also compared participants’ attitudes toward people with different opinions using self-rated receptiveness scores before and after using the Baseline system and FrameIt. On the *Single-use Plastic Ban* issue, participants reported a significant increase in the perceived informativeness of opposing opinions (Wilcoxon Signed-Rank test, $p = 0.016$, $W = 0.0$) and curiosity on opinion difference (Wilcoxon Signed-Rank test, $p = 0.046$, $W = 6.0$) after using FrameIt. In contrast, none of the measures showed significant pre-post differences when using the baseline (Table 5.7). On the *Cancel Culture* issue, willingness to converse with people with different opinions increased significantly after using FrameIt (Wilcoxon Signed-Rank test, $p = 0.011$, $W = 0.0$) (Table 5.8). We attribute these topic-wise differences in willingness to converse as well as informativeness to the underlying characteristics of the topics. *Cancel Culture* was perceived as more strongly related to ethical concerns than to factual content, which may explain why participants benefited more from FrameIt in this context. In contrast, participants regarded the *Single-Use Plastic Ban* issue as involving both factual information and moral values, which might have led to improvement in perceived informativeness.

5.7 User Study 2: Deployment-based User Study

Contrary to the comparative lab study described in Sections 5.5 and 5.6, we conducted a five-day deployment study to further understand how users engage with one another through FrameIt. The goal of the study was to show an ideal environment where everyone uses FrameIt, with the goal of observing the maximal effects of using FrameIt. Not having a baseline was a deliberate choice, as the goal of the study was to observe a discussion in a homogeneous environment using only FrameIt.

Table 5.7: Pre-post difference of self-evaluated receptiveness of the participants on Single-use Plastic Ban issue. While there was no significant difference in the baseline condition, participants in FrameIt condition had positive changes in the perceived informativeness and curiosity to different opinions.

	Baseline			FrameIt		
	Pre	Post	p / W	Pre	Post	p / W
On this issue, I am willing to have conversations with individuals who hold strong views opposite to my own.	5.55	5.18	0.16 / 2.5	5.55	5.82	0.46 / 7.0
On this issue, I like reading well thought-out information and arguments supporting viewpoints opposite to mine.	5.90	5.54	0.36 / 2.5	5.55	5.64	1.0 / 7.5
I find listening to opposing views on this issue informative.	5.45	5.90	0.42 / 4.5	4.82	6.0	0.016 / 0.0*
I value interactions with people who hold strong views opposite to mine on this issue.	5.09	5.54	0.29 / 10.5	4.82	5.28	0.52 / 7.5
On this issue, I am generally curious to find out why other people have different opinions than I do.	5.45	5.54	0.65 / 6.0	4.64	5.55	0.046 / 6.0*
I feel that people who have opinions that are opposite to mine on this issue often have views which are too extreme to be taken seriously.	4.36	4.36	0.93 / 13.5	4.18	4.55	0.51 / 17.0
On this issue, I feel that people who have views that oppose mine rarely present compelling arguments.	3.91	4.45	0.12 / 5.0	4.82	3.91	0.23 / 12.5
I feel that information regarding this issue from people who have strong opinions that oppose mine is often designed to mislead less-informed listeners.	4.82	4.73	0.79 / 12.5	4.18	3.82	0.55 / 12.5
I feel that people who have views that oppose mine on this issue are often biased by what would be best for them and their group.	5.0	4.64	0.31 / 11.0	4.91	4.09	0.14 / 3.5
I feel that people who have views that oppose mine on this issue often base their arguments on emotion rather than logic.	4.64	4.19	0.47 / 16.5	3.91	3.28	0.11 / 7.0

Table 5.8: Pre-post difference of self-evaluated receptiveness of the participants on Cancel Culture issue. While there was no significant difference in the baseline condition, participants in FrameIt condition had positive changes in the willingness to have conversation with people with opposing views.

	Baseline			FrameIt		
	Pre	Post	p / W	Pre	Post	p / W
On this issue, I am willing to have conversations with individuals who hold strong views opposite to my own.	5.64	5.81	0.49 / 5.0	4.55	5.27	0.011 / 0.0*
On this issue, I like reading well thought-out information and arguments supporting viewpoints opposite to mine.	5.81	5.81	0.73 / 12.0	5.55	6.0	0.096 / 3.0
I find listening to opposing views on this issue informative.	5.00	4.64	0.42 / 16.0	5.63	5.28	0.19 / 1.5
I value interactions with people who hold strong views opposite to mine on this issue.	5.00	4.82	0.68 / 6.0	5.18	5.45	0.52 / 13.5
On this issue, I am generally curious to find out why other people have different opinions than I do.	5.72	5.45	0.57 / 14.0	5.91	5.82	0.74 / 9.0
I feel that people who have opinions that are opposite to mine on this issue often have views which are too extreme to be taken seriously.	3.63	4.09	0.21 / 7.0	5.09	4.09	0.10 / 9.0
On this issue, I feel that people who have views that oppose mine rarely present compelling arguments.	3.73	4.09	0.41 / 4.5	3.73	4.0	0.73 / 12.0
I feel that information regarding this issue from people who have strong opinions that oppose mine is often designed to mislead less-informed listeners.	4.00	4.72	0.13 / 2.0	4.18	4.18	0.48 / 10.0
I feel that people who have views that oppose mine on this issue are often biased by what would be best for them and their group.	4.09	4.36	0.59 / 18.0	4.82	4.45	0.60 / 11.0
I feel that people who have views that oppose mine on this issue often base their arguments on emotion rather than logic.	4.27	4.27	0.86 / 13.0	4.73	4.36	0.44 / 9.5

Table 5.9: Demographic profiles of the field study participants and their prior opinions on the three issues used in the experiment. The prior opinions were measured in 7-point Likert scale (1: Strongly Disagree, 7: Strongly Agree).

No.	Age	Gender	Nationality	Prior opinion (1–7 Likert)		
				Plastic Ban	Animal Testing	Progressive Fining
DP1	25	Male	South Korea	5	6	3
DP2	22	Female	Kazakhstan	5	7	3
DP3	22	Male	Ethiopia	6	5	7
DP4	25	Male	South Korea	3	2	6
DP5	18	Female	South Korea	3	5	2
DP6	20	Female	Kazakhstan	3	5	4
DP7	30	Male	South Korea	2	2	4
DP8	26	Male	South Korea	3	1	5
DP9	26	Male	South Korea	3	5	2
DP10	24	Female	South Korea	5	1	5
DP11	21	Male	South Korea	6	5	6
DP12	21	Female	Philippines	5	2	6
DP13	22	Male	Kyrgyzstan	7	7	5
DP14	22	Female	Vietnam	1	5	7
DP15	18	Female	South Korea	3	3	6

5.7.1 Participants

We recruited 15 participants from an engineering university in South Korea. Our participants (8 male, 7 female) were aged 18 to 30 (Median = 22) and had various nationalities. Table 5.9 summarizes their profiles, including age, gender, nationality, and prior opinions on the three issues discussed during the study. Each participant was compensated with KRW 50,000 (approximately USD 36). This study received approval from a university’s Institutional Review Board (IRB).

5.7.2 Materials

Unlike the first user study, the deployment study involved a single condition, FrameIt, to qualitatively examine its impact on participants’ perceptions of others and their comment-writing process in real online discussions, by creating a homogeneous discussion environment to observe the full effects.

To broaden the scope of analysis, we retained one topic from the previous study and introduced two new topics. For each topic, we provided a neutral description of the topic along with commonly raised arguments, adopted from existing discussion forums, to serve as a discussion starter:

- Should disposable plastic items be banned?⁵
- Should animal testing be banned?⁶
- Should fines be proportional to income (progressive fining)?⁷

⁵<https://www.kialo.com/disposable-plastic-items-should-be-banned-29627>

⁶<https://www.kialo.com/should-animal-testing-be-banned-37070>

⁷<https://introductiontoethics.pressbooks.tru.ca/chapter/ethics-bowl-case-progressive-fines>

5.7.3 Procedure

The study ran for five days, excluding the pre-study tasks and the post-study interview.

Pre-study Tasks Prior to the first day, participants completed 7-point Likert Scale questions to indicate their stance on the discussion topics. They also received a slide deck introducing the Moral Foundations Theory and FrameIt, identical to the materials used in the comparative lab study (Section 5.5). To support participants in better understanding the Moral Foundations Theory, we asked them to test their own moral foundations priorities by completing the Moral Foundations Theory Questionnaire⁸.

After the pre-study tasks, participants engaged in the following activities on each respective day:

- **Day 1:** The participants wrote **one** comment on their assigned topic. The participants were randomly assigned.
- **Day 2:** The participants picked a topic and wrote **at least two** comments in reply to the comments of the others from Day 1.
- **Day 3-4:** The participants first revisited their comments from previous days. If they had replies from other participants, they replied to at least one of them. Then, the participants selected a topic that they had not participated in during the previous days and wrote **at least one** comment as a reply to the comments of the others. If they had not received any reply to their comment, they wrote **at least two** comments to others' comments.
- **Day 5:** The participants first revisited their comments from previous days. If they had received replies from other participants, they replied to at least one of them. If they had not received any reply to their comment, they wrote **at least one** comment to others' comments.

Throughout the study, participants used aliases to participate in discussions. To ensure sincere participation, participants were asked to write more than 100 words for each comment and were instructed not to use external LLM tools when composing their comments.

By the end of Day 5, all participants had contributed to each of the three discussion topics. Participation was considered incomplete if a participant failed to complete the task on any day or did not engage with all three discussion topics.

After completing the tasks for 5 consecutive days, participants were asked to complete a post-task survey consisting of Likert-scale questions about their impressions of FrameIt and their opinions on the discussion topics.

After completing the post-task survey, participants were asked to participate in a one-hour semi-structured interview. During the interview, participants reflected on their experience of interacting with others through FrameIt, discussed each system feature, and shared holistic impressions, including perceived discussion quality and a sense of ownership.

5.8 User Study 2 Results

In this section, we present qualitative findings on how participants perceived the discussion experience with FrameIt and how FrameIt influenced their interactions with others.

⁸<https://moralfoundations.org/questionnaires/>

5.8.1 Change in Discussion Atmosphere

Participants generally perceived discussions empowered by FrameIt as more respectful than typical online discussions. They described conventional online discussions as “hostile to people with different opinions” (DP5) and “a risk of being attacked” (DP11), which prevents them from actively participating. In contrast, they perceived discussions with FrameIt as more formal, polite, and respectful, as the comments in FrameIt often acknowledged opposing views and sought to complement rather than aggressively refute them. DP9 described that such an atmosphere encouraged them to express respect for others’ opinions as well.

5.8.2 Scaffolding Respectful Communication

FrameIt encouraged participants to frame their discussions through the lens of moral foundations. In this section, we present how participants used FrameIt’s features and illustrate how these features shaped their experiences in real discussion settings.

Moral Foundations as Cues for Discovering Different Perspectives

Some participants (DP14) mentioned that they tried to discover opinions different from their own. In that case, differences in moral foundations were often perceived as indicators of distinct stances on a given issue, creating space for participants to discover new perspectives and arguments on the topic. In some cases, participants became aware of different interpretations of the same moral foundation, which also helped respect different opinions. DP4 reflected: “I advocated for the Progressive Fine, believing that everyone should have an equal weight of financial burden. Yet, some criticized this as unfair. It came down to a difference in how we value fairness: equitable burden versus absolute numerical equality.”

Moral Foundations as Cues for Careful Reading

The tags of moral foundations and sentence-level annotations were often used as an overview of arguments guiding how to read the comments. Some participants valued them as a way to “grasp the main points” (DP5), focusing on the annotated sentences (DP13). Some other participants thought that the annotations encouraged them to read the comments more thoroughly. For instance, DP9 and DP14 perceived comments with more moral foundation labels as more comprehensive about the issue, which encouraged them to read the comment in more detail. Participants perceived such guidance as helpful for thoroughly understanding others’ comments. DP7 said, “For the single-use plastic ban, for instance, I noticed that ‘Care’ and ‘Authority’ foundations were very prominent, rather than ‘Fairness.’ These were values I hadn’t usually prioritized on myself, so I had to be mindful. While reading others’ comments, I constantly asked myself, ‘What does Authority mean to this person?’ or ‘What does Care mean to them?’ ”.

Perspective-taking

When the participants read comments with different stances on the issues, the moral foundations helped them better understand the opinions and underlying arguments. Participants (DP3, DP4, DP5, DP6, DP10, DP12) noted that the explicit labels for moral foundations helped them better understand why people held different opinions. DP12 explained, “With the (Moral Foundation) labels, I got to know about the users more and about their values”. By being aware of such differences, participants (DP1,

DP3, DP4, DP5, DP7, DP10) felt they could be more open to different opinions. DP4 said, “If the moral foundations someone else uses are different from mine, it sets a baseline that we are looking at things from different angles. Knowing that there will be differences before I even start reading allows me to be more open-minded and tolerant of their opinion.”

Deepening Thoughts via Aligning Moral Foundations

With explicit labels for moral foundations and FrameIt’s support for moral alignment, participants felt they should write comments while considering others’ moral foundations. DP5 explained, “When leaving a comment, I took into account the moral foundations reflected in the previous user’s comment. This led me to formulate my own input as a supplement to theirs, either by offering counterarguments supported by different evidence or by expressing agreement.” Engaging with such comments not only broadened their understanding but also encouraged them to express their own stance through unfamiliar moral foundations, thereby pushing the boundaries of their thoughts and perception on a topic. DP4 said, “When I encounter themes that others have used which I hadn’t considered, it forces me to keep reflecting, which ultimately helps expand my own thinking.” While writing comments from unfamiliar moral foundations, the *Suggest Discussion Points* feature served as a support to expand their reasoning, particularly on topics where they held weak prior opinions or felt uncertain about how to substantiate their stance on the issue. DP4 said, “When I have debates or disagreements with friends, sometimes they point out things I would’ve never thought of. Suggested discussion points did the same—highlighting those spots in advance, which helped me focus during the discussion, prepare counterarguments, and expand my ideas.” Furthermore, DP4 and DP7 described that the overall process of writing comments while aligning with the moral foundations helped them become more aware of their own moral foundations.

Facilitating Reflection

Participants felt that discussing with FrameIt prompted them to reflect on the quality and the respectfulness of their comments. DP9 said, “Normally, when I write comments in online communities, I don’t think about the other person’s feelings—I just ‘dump’ my own thoughts. But here, I felt more pressure because I had to consider the other person’s feelings.” Participants used the *Generate Feedback* feature as the final step before clicking the submit button. They perceived the feature as a way to gauge how others might perceive their comment, especially its weaknesses, before being refuted by other discussion participants. DP10 said, “I felt that the feedback was a way to check how structured and persuasive my opinion was.” The real-time *Moral Alignment Indicator* also encouraged the participants to reflect on their tone. DP7 shared their experience, “The real-time indicator helped me not to think too emotionally and subjectively. I sometimes leave very subjective comments, but the features of the system kept encouraging me to consider the moral foundations.”

Facilitating Agreeing to Disagree

Participants often regarded the differences in moral foundations as inherent characteristics of individuals rather than as positions open to change. It led some participants (DP3, DP5, DP7, DP10, DP15) to perceive opinion differences as stemming from fundamental value differences. In such cases, they chose to “agree to disagree”, rather than engage in overheated discussions to change others’ opinions. For instance, DP10 explained: “In discussions, I usually try to persuade others or gain sympathy for my view.

But when I saw someone with completely different moral values, I thought their values wouldn't change. So I simply wrote, 'This is my position, and that is yours.' On the positive side, DP3, DP5, and DP7 thought such an atmosphere could help people be more open to conversations with others with different opinions to better understand each other, rather than engage in overheated discussions. On the other hand, DP5 and DP10 thought that "agreeing to disagree" might not help participants reach a shared conclusion.

5.8.3 Tensions

FrameIt was not always perceived positively.

Blurring the Contestation

Some participants had mixed opinions on whether the discussion atmosphere with FrameIt was desirable. DP8 and DP11 argued that productive online discussions should surface contestation between different opinions, enabling the participants to think on the relative weights and societal significance of opposing opinions. However, DP8 thought that such contestations were blurred while other participants were trying to respect others' opinions, leading to "middle-ground" comments that, while rational, failed to engage them in contestation. The experience of DP7 showed an instance: "It was fascinating to see the root of our different perspectives. But because we understood each other's positions so clearly, there wasn't a 'clash' of opinions. In fact, knowing exactly where they were coming from made it harder to keep the discussion going."

Excessive Cognitive Load

DP8, DP9, DP10, and DP13 thought the discussion on FrameIt was cognitively demanding and time-consuming, limiting FrameIt's application to real-world online discussions. For instance, DP1, DP7, DP8, DP11, and DP12 felt that the *Generate Feedback* feature often demanded too rigorous arguments for "classic" online discussions. DP8 remarked, "The feedback often pointed out the error of overgeneralization. I was aware of it, but since the purpose of the discussion was to debate what's more important, I tended to argue more strongly. Others probably followed the feedback more. I think that's why others wrote softer comments to avoid being attacked from either side."

At the same time, participants emphasized that the design of the *Generate Feedback* feature could be more specific and actionable. For example, DP8 remarked, "When it comes to feedback, it usually just points out what's wrong. I think it'd be nice if it also gave me follow-up questions based on my writing — like questions that challenge or respond to what I wrote. Instead of just a block of text, a list of questions would feel more like a real conversation and be easier to follow."

Experiencing Feedback as Supportive yet Intrusive

Some participants perceived the argumentation support features of FrameIt as overly intrusive, "taking the fun out of writing their own ideas" (DP11). For instance, DP11 felt that *Moral Alignment Indicator* prevented them from expressing their own views, as they felt reluctant to share their thoughts when FrameIt encouraged them to adopt different moral foundations in their argument. With *Suggest Discussion Points* and *Generate Feedback* features, participants (DP3, DP11) felt the generated points did not feel like their own writing. DP2 said they often adopted the FrameIt-generated discussion points

when they were stronger than DP2’s own points, but at the same time, they felt their comment was ‘artificial’.

Trust on the Analysis of FrameIt

The participants considered that the moral foundations tagged on the comment and each sentence were acceptable in general, at least for their own comments. For instance, DP1 said, “I wrote a comment around ‘Fairness/Cheating’ foundation. However, the system annotated one sentence with ‘Care/Harm’.

While it technically made sense, I felt it could disrupt the overall flow of the argument. Other users might not notice it if they did not hover on the tag.” DP4 said, “I think the annotations were accurate in general. Also, many sentences can be interpreted in multiple ways, so I could understand why it chose those specific labels.”

5.9 Discussion

5.9.1 Theory of Change

In this paper, we designed FrameIt to foster “respectful engagement” among people with differing opinions. By externalizing others’ moral foundations, we aimed to help participants better understand the motivations and reasoning behind others’ stances, and to shift users’ attributions away from assumptions of bad intent or irrationality toward legitimate differences in values.

Our findings from the user studies showed that surfacing moral foundations indeed helped participants better understand others, contributing to a respectful discussion environment. Findings from our comparative user study demonstrated that awareness of others’ moral foundations helped participants better understand others’ opinions and adopt a more receptive attitude towards them. Our 5-day deployment study provided qualitative evidence for the mechanism underlying this difference in attitude: awareness of others’ moral foundations facilitated perspective-taking, enabling better understanding of the reasoning behind people’s opinions. Furthermore, FrameIt’s support for writing opinions aligned with the moral foundations further encouraged participants to think more deeply about the issues and reflect on their opinions. This pattern is consistent with the attribution shift we hypothesized: participants engaged with a disagreement as a difference in what others valued, rather than as a sign of the other person’s ignorance or bad faith.

At the same time, our receptiveness measures are more consistent with a *recognition* account than with a reduction in *derogation*. Across both topics, the items that moved were those concerning the perceived value of engaging with opposing views: informativeness and curiosity on the Single-Use Plastic Ban (Table 5.7) and willingness to converse on Cancel Culture (Table 5.8). Items expressing derogation of opponents, such as rating opposing views as “too extreme,” “biased,” or “based on emotion rather than logic,” often trended in the expected direction but did not reach significance. This suggests that the clearest and most consistent shifts were in how *intelligible and worth engaging with* participants found opposing views, rather than in how negatively they judged those who held them. This suggests that FrameIt’s effect may be narrower than our initial hypothesis: surfacing value differences supported recognition of opposing views as coherent and informative, which is a precondition for respectful engagement, but we did not find clear evidence that it dissolved negative judgments of opponents.

While prior research has shown that providing diverse contextual information around comments can shape readers’ attitudes [119, 81, 210, 174, 79, 80], our findings extend this line of work by demonstrating

that making implicit moral foundations explicit can also positively influence how readers perceive and engage in online discussions. However, we only observed these effects over a short period of time. Future work should therefore examine whether surfacing moral foundations leads to sustained attitudinal and behavioral changes through larger-scale longitudinal deployment studies.

5.9.2 Agreeing to Disagree: Respect at the Cost of Contestation

When participants recognized a disagreement as a difference in underlying values, they often treated that difference as a fixed trait of the person rather than a position open to change, and chose to “agree to disagree”. This had a clear upside. Rather than dismissing an opposing view, participants (DP3, DP5, DP7) accepted it as a legitimate position and avoided overheated argument, which is the respectful engagement we set out to support. But the same acceptance limited deeper engagement. Some participants (DP5, DP10) noted that “agreeing to disagree” left them without a shared conclusion. Others questioned whether a respectful exchange should be the goal at all: DP8 felt the shared drive to respect opposing views flattened the discussion into rational but inert “middle-ground” comments, and DP11 argued that productive discussion should instead surface and weigh competing positions. These two outcomes may stem from the same underlying shift: seeing an opposing position as grounded in a stable value makes it worth respecting and also makes it look unchangeable.

These outcomes reflect a tension between two ideals of discourse. We designed FrameIt around a deliberative ideal: a respectful exchange of logically concrete arguments (DG1, DG2), in line with the norms of deliberative democracy [37]. Its features were designed to support reflection on others’ moral values and to improve the logical quality of comments. Some participants, however, valued an agonistic ideal, in which open contestation between opposing positions is itself what drives democratic discourse [34, 211]. For them, understanding why others held a position made the discussion easier to respect but harder to continue. This suggests that the deliberative ideal we assumed is not universal, and that the value of surfacing moral foundations depends on what a discussion is for.

FrameIt should therefore be tailored to the discussion’s purpose. When the goal is consensus, such as collaborative decision-making, accelerating mutual understanding is useful. When the goal is contestation, the same interactions could be applied in different ways. The system could surface shared foundations first to prevent early disengagement and reveal differences only once a discussion begins to overheat. It could also be selectively activated, for instance, to foster respect for minority viewpoints rather than to smooth over every disagreement.

5.9.3 Conceptual Generalization Beyond Moral Foundations

While FrameIt leverages Moral Foundations Theory (MFT) as its primary interpretive lens, we believe that our findings point to a more general design pattern: an interactive system can mediate disagreement by making implicit value structures explicit during interaction. The core mechanism is not the specific taxonomy, but the introduction of a shared interpretive scaffold that redirects attention from surface-level claims to the underlying priorities they express. Consistent with this reading, our results represent a successful application of one particular value framework rather than evidence for the universal adequacy of MFT.

We selected MFT because it offers an empirically grounded and computationally tractable account of the moral pluralism behind opinions on social and political issues. However, many disagreements are only partly moral. For instance, disagreements about public health policy during a pandemic may be partially

explained through moral foundations [196, 212], but they also reflect trade-offs among economic stability, public safety, individual liberty, and cultural practices [213]. These dimensions extend beyond moral reasoning. Existing computational approaches to issue framing and value detection could therefore be integrated to surface non-moral priorities, broadening the scaffold while preserving the same underlying mechanism.

Generalizing beyond MFT raises important research questions. Different value frameworks vary in granularity, cultural specificity, and normative assumptions. The choice of taxonomy may influence not only how disagreement is interpreted, but also how it is transformed through interaction. We observed two effects with MFT: participants better recognized opposing positions as coherent, but also came to see value differences as fixed. Whether these same effects occur with other value frameworks is an open question, and future work could test this by directly comparing frameworks.

5.9.4 Ethical Implications

FrameIt algorithmically interprets and surfaces the moral structure of user-generated content. This leads to a sensitive normative position: it does not merely facilitate discussion, but actively mediates how disagreement is framed and understood. We discuss several ethical considerations from our positioning:

Over-reliance on Interpretive Labels Our technical evaluation showed that the pipeline deliberately favors recall over precision, over-annotating moral foundations relative to human annotators (Section 5.4). Yet participants generally accepted the system’s labels as reasonable. The combination of imperfect, over-assigned labels and high user trust may enable FrameIt to quietly shape judgments before users form their own. In our studies, we observed instances in which participants accepted the system’s analysis without critically evaluating its validity, suggesting the risk of over-reliance. This dynamic is particularly concerning given that moral foundation attribution is inherently subjective and pluralistic; previous work suggests that even human annotators often disagree on moral categorizations [202, 214, 198]. When users defer to AI-generated moral labels, the system may unintentionally narrow interpretive diversity and present contested classifications as if they were objective facts. Designing such systems, therefore, requires careful preservation of user agency. AI-generated frames should be presented as hypotheses rather than authoritative judgments, and interfaces should explicitly communicate uncertainty and encourage critical reflection rather than passive acceptance.

Bias and Cultural Partiality in Moral Interpretation The LLMs powering FrameIt are trained on datasets that potentially reflect historical and cultural imbalances. Prior work has shown that such models disproportionately encode Western-centric perspectives and may amplify political or cultural biases present in their training data [215, 216]. This raises particular concerns, as FrameIt provides interpretations of moral foundations. Prior research suggests that Western and politically liberal populations tend to emphasize the Care/Harm and Fairness/Cheating foundations more strongly than others [187, 217]. If the underlying models share or amplify these distributions, FrameIt may systematically over-detect certain foundations while under-representing others, subtly shaping discourse in culturally partial ways. Such asymmetries are especially consequential in political or cross-cultural discussions, where differential sensitivity to particular moral foundations may influence how disagreement is interpreted. We believe that deploying FrameIt would require ongoing auditing, cross-cultural evaluation, and calibration to mitigate systematic skew.

Manipulation and Strategic Moral Framing Making moral foundations explicit may also enable manipulation of public opinions. While surfacing moral structure and arguing in terms of others’ moral foundations can support empathy and reflection, it can also be leveraged to craft persuasive or manipulative messaging tailored to others’ moral profiles. In adversarial contexts, such affordances could be used to reinforce misinformation, deepen polarization, or exploit value-based vulnerabilities through different cultural or political contexts. Mitigating such misuse requires careful consideration of access controls, transparency about system decisions and reasonings, and constraints on how moral analyses are surfaced or operationalized.

5.10 Limitations and Future Work

We recognize several limitations of our work. First, we tested FrameIt with a limited group of participants recruited from universities. In both studies, participants were generally well educated, suggesting they found it easier to understand the concept of moral foundations and to write well-supported arguments. Also, although the participants were fluent in English, English was not their primary language, which could have affected their commenting behavior. Moreover, in both studies, interview responses indicated that some participants exhibited socially desirable behaviors, such as greater respectfulness and politeness, due to the study’s setup.

Second, we acknowledge that the more recent versions of the MFT with more fine-grained sets of values, such as including Liberty/Oppression or dividing Fairness/Cheating into Equality and Proportionality, may better explain people’s moral reasoning. Despite that, we chose the initial version of the MFT because it was the most widely used, not only in communication and psychology but also in natural language processing and computational social science.

Future work could address these limitations through a deployment study of FrameIt on diverse social media and discussion forums, for instance, by developing it as an add-on. Such a deployment study could also enable us to investigate the generalizability of FrameIt, including how replies differ when moral foundations are considered. This could give insight into potential changes in the discussion tone, constructiveness, quality, and overall atmosphere.

5.11 Conclusion

Constructive online discussions can, under the right circumstances, promote mutual understanding across different opinions and foster common ground. However, in practice, many online discussions are characterized by incivility and disrespect, which hinders meaningful exchange of opinions between participants. To address this challenge, we built FrameIt, an LLM-powered system for respectful and constructive discussion that explicitly surfaces participants’ value systems grounded in the Moral Foundations Theory. FrameIt identifies moral foundations in other participants’ comments using a set of LLM prompts and supports users in writing comments aligned with these moral foundations. FrameIt provides three forms of support: (1) real-time analysis of alignment of moral foundations, (2) suggestions for discussion points grounded in others’ moral foundations, and (3) feedback on comments in terms of both logical quality and moral alignment. From our user evaluations, we found that FrameIt helped participants recognize the moral values of others’ perspectives, enabling them to engage more openly with differing opinions and contribute more constructive and respectful comments. At the same time, we observed a tension between “agreeing to disagree” with different opinions, adopting a respectful attitude,

and hindering contestation between conflicting viewpoints. Building on these insights, we discuss design implications for future AI-mediated discussion systems that abstract and mediate between differing perspectives on complex issues. We expect that the findings and design lessons from this work will inform future research on fostering more constructive and respectful online discussions, ultimately contributing to a more civil, respectful, constructive, and inclusive online discourse environment.

Chapter 6. DecideTogether: Supporting Automoderation with Subjective Criteria through Human-AI Collaborative Calibration

This chapter presents DecideTogether. All uses of “we”, “our”, and “us” in this chapter refer to the collaborators who contributed to this work.

6.1 Motivation & Contributions

Online communities rely on rules to govern member behaviors and sustain their distinct identities, norms, and values [105, 218]. Although communities often face structurally similar challenges, such as spam, personal attacks, off-topic discussion, and low-quality contributions, each community develops its own rules that reflect its particular context [219, 220, 218, 221]. A “Be Kind” rule, for example, can point to a tone that one community treats as baseline civility and another treats as overly restrictive; what counts as “frank advice” in a career-focused community may read as gatekeeping in another. The same behavior can be acceptable in one community and a violation in another. The answer to whether a given post belongs is community-specific and case-specific [222]. When moderators develop and enforce such rules, they exercise discretion shaped by an evolving sense of what and who the community is for [220, 221].

When a moderator creates a rule, this *moderation intent*, the moderator’s sense of what behavior fits the community, manifests in two forms. The first is the *rule text*, which communicates the stance to community members and sets shared expectations about what is acceptable. The second is the *automoderator logic*, which enforces the stance at scale in a machine-executable form such as regular expressions or keyword lists [223, 224, 225]. Because human moderators cannot review every submission at the scale of active communities, most platforms rely on tools such as Reddit’s automoderator [226] and Discord’s MEE6 [227] to implement this logic-based enforcement. They work for rules whose violations are lexically signaled, such as a banned slur, a forbidden URL, or an off-topic flair, but are poorly suited to interpretive rules whose violations depend on tone, context, and the moderator’s reading of intent rather than on any specific word [228, 229, 230, 222]. Each form expresses the same moderation intent in a different register, and the moderator’s task is to articulate that intent into both.

Carrying intent into each form is difficult, and for different reasons. The first difficulty concerns the rule text. Moderators’ enforcement intent is not fully formed in advance but is articulated through iteration [231]. What a rule like “Be Kind” actually filters out in a given community often becomes legible only when moderators are confronted with borderline cases within their own community and make decisions about them. The intent is tacit, grounded in norms and tone that resist direct statement, so committing it to rule text is itself an act of articulation rather than transcription.

The second difficulty concerns the automoderator logic. Even when a moderator has a settled sense of intent, translating it into consistent automated behavior is hard. Data-driven classifiers trained on past moderation decisions address some subjective cases, but they often lack explainability and require large amounts of data to provide community-specific moderation [232, 233]. Recent systems instead use LLMs as judges that assess whether a post violates a written rule [225, 224, 234, 235], but to enforce a

subjective rule, the model must fill in what the rule leaves unsaid, and its reading can diverge from the moderator’s. Because the logic runs at scale, a gap between the intended and the enforced reading can go unnoticed. The LLM’s stochasticity compounds this, since the same post can be judged differently across runs. This opacity is a substantive obstacle for a high-stakes, accountable activity: moderators need to see why the system reached a given decision in order to trust it and to correct it [224]. Closing the gap between intent and system behavior thus requires iteration, running the system on real posts, evaluating its decisions, and revising the prompt or rule. Yet moderators currently face a steep gulf of execution and evaluation in this loop, since it is rarely clear which edit to the prompt will produce the desired change in behavior, or whether an observed change reflects a genuine fix or stochastic noise [236, 224].

In this chapter, we present DecideTogether, an interactive system that supports moderators articulating their community-specific subjective moderation intent into community rules and automoderator logics. It supports moderators in articulating their community-specific subjective intent into a rule text, and in translating that intent into automoderator logic whose behavior they can inspect and align. By sampling posts and comments, DecideTogether first provides a community profile of the community’s norms, represented by predefined tags across four dimensions: Purpose, Participants, Stakes, and Tone. When the user writes a rule to describe their moderation intent, DecideTogether provides LLM-generated rule text suggestions based on the rule’s high-level goal and the community profile, to support incorporating context-dependent rationale into the rule. Based on the rule text, DecideTogether produces a set of binary questions that serve as the internal logic for automoderation, designed to ensure the consistency of the LLM’s verdicts and to provide fine-grained explanations of the decisions. The user can test the alignment between the automoderation behavior of DecideTogether and their moderation goal and criteria by examining the automoderator-generated verdicts on the real community posts and comments in the moderation queue. The user can further refine their moderation intent by editing the rule texts, specifying relevant community profile tags around the rule, or editing the logic directly. DecideTogether supports the refining process by providing live previews of how the change will impact moderation decisions and by suggesting updates to the logic based on previous error cases.

We evaluated DecideTogether with a technical evaluation and a user study. With the technical evaluation, we evaluated whether DecideTogether’s rule-compilation pipeline preserves the goals and intent of natural-language rules. Using an LLM-as-a-judge approach, we evaluated the quality of compiled logic across four dimensions: coverage, logical correctness, minimality, and clarity, comparing logic compiled with and without community context. Furthermore, we evaluated moderation-decision accuracy on a ground-truth moderation dataset constructed from 15 subreddits spanning Q&A, advice, debate, and hobby communities, and examined whether iteratively updating the compiled logic from observed errors improves alignment with human decisions. Our results showed that the automoderator logic generated by DecideTogether accurately captured the goals and intents of the rule text and substantially outperformed direct rule-text prompting in replicating human moderators’ decisions, reaching 65.9% accuracy after error-driven updates. Adding community context produced logic better aligned with moderation intent, though its effect on decision accuracy varied with the granularity of a community’s existing rules.

To explore how DecideTogether supports moderators in shaping their moderation goals and criteria, and in aligning the automoderator with their intent, we conducted a qualitative user study with 10 community moderators. To thoroughly observe the process of shaping the moderation goals, we asked the participants to create a new rule to moderate problematic posts involving subjective decision-making for a hypothetical community. The findings from our qualitative observations and interviews showed that DecideTogether enabled its users to articulate and refine their moderation intents through a bidirectional

dialogue, in which they examined how DecideTogether interpreted their initially vague or tacit intents into automoderator logics and moderation verdicts on posts, and revised their intents in response. Beyond intent articulation, we found that the subjective nature of DecideTogether’s decisions reshaped how moderators perceived trust and control: rather than treating control as a parallel concern, participants regarded it as a precondition for trust, finding DecideTogether’s contestable subjective decisions trustworthy precisely because they were paired with non-autonomous execution that deferred final judgment to human moderators.

This chapter provides the following contributions:

- DecideTogether, an interactive system that scaffolds moderation intent into both a communicable rule text and an inspectable, alignable automoderator logic within a community context
- A technical evaluation showing that DecideTogether’s rule-compilation pipeline preserves the intent of natural-language rules and that error-driven logic updates improve alignment with human moderation decisions
- Empirical findings showing how DecideTogether supports expression and refinement of moderation intents and how DecideTogether shapes moderators’ notions of trust and control for subjective automoderation

6.2 Design Goals

In this section, we present design goals for supporting moderators to articulate their intents for automating subjective moderation. We ground these goals in established findings from prior work on automoderation and on aligning LLM behavior with human intent.

6.2.1 DG1: Surface Latent Criteria in Community Context

Moderators begin from a high-level intent whose operative criteria are not yet explicit. What a rule, such as “Be Kind”, should filter is underspecified at the outset and depends on the community’s norms, values, and recurring cases, which makes the same words mean different things across communities [222, 219, 237, 218]. Because the intent is tacit and grounded in community-specific expectations, moderators struggle to spell out these criteria directly when authoring rule text or configuring automoderation [226, 230].

The system should help moderators surface the latent criteria implied by their high-level intent, interpreted in the context of the community rather than the rule string alone. Rather than asking the moderator to specify an exhaustive rule in advance, the system should make explicit the criteria for the intent that this community entails, so that articulation begins from a concrete reading of what the rule means here.

6.2.2 DG2: Make Automoderation Results Usable for Intent Articulation

Automoderation results are a moderator’s main evidence about whether a rule captures their intent, but current tools make them hard to read for this purpose. Regex-based tools fail in ways close to debugging: a result reflects whether the pattern matched, not whether the intent was right [226]. LLM-based moderation introduces the opposite problem, since the model fills in what the rule leaves unsaid

and its output varies across runs, so a wrong decision is ambiguous between a misstated intent and model noise [225, 224, 235]. When an error cannot be attributed, the result cannot inform articulation.

The system should make automoderation results usable as evidence about intent. This requires reducing incidental variation so that a change in output reflects a change in the rule rather than stochastic noise, and presenting decisions so that a moderator can attribute a mismatch to their own intent and revise it. The aim is to turn results from outcomes to be debugged into signals that guide articulation.

6.2.3 DG3: Keep Rule Authoring and Revision Legible Without Technical Expertise

Moderators often lack the technical background to author or adjust automoderation logic directly, and regex-era tooling imposed a high cost on every change [226]. Lowering the surface cost of editing is not sufficient on its own: when a wording change propagates into machine behavior opaquely, moderators lose track of what their rule now does [236].

The system should let moderators author and revise rules in natural language while keeping the effect of each revision legible, showing how a change to the rule alters concrete decisions. Editing should not require operating on the logic directly; instead, revising the rule in accessible terms should be the means through which the underlying logic is updated, with the consequences of that update made visible.

6.2.4 DG4: Make Moderation Decisions Transparent and Traceable

Subjective moderation decisions are contestable: community members may disagree with them and demand justification, so the legitimacy of a decision depends on being able to show how it was reached [238, 239]. In a high-stakes, accountable activity, a verdict that cannot be explained cannot be defended when challenged.

The system should make each decision transparent and traceable, exposing the reasoning that led to a verdict and linking it back to the criterion and rule text that produced it. Bi-directional traceability among rule text, decision logic, and verdicts lets a moderator move from a decision to the criterion behind it and to the rule text that shaped it, supporting the accountability that contestable decisions require.

6.3 System

In this section, we introduce the design and implementation details of DecideTogether.

6.3.1 Key Idea: Aligning Rule Text and Automoderator Logic

A moderation intent must take two forms to be enacted. It must be communicated to community members as a rule that states what the community expects, and it must be executed by a machine as logic that decides individual cases. DecideTogether treats these as two expressions of a single intent: a community-facing rule text and a machine-facing automoderator logic. The rule text is written in natural language for community members to read. The automoderator logic is the executable criteria the system applies to each post and comment.

To keep the logic inspectable, DecideTogether does not treat it as a single opaque prompt. Instead, it decomposes the rule into a set of binary questions, each pairing a criterion with an action taken when the

criterion is met. These criteria range from objective ones, such as a banned keyword or a structural signal, to subjective ones, such as whether a comment is dismissive in tone, which the system evaluates with an LLM rather than a fixed pattern. Decomposing even the subjective criteria into discrete questions, rather than delegating the entire judgment to a single prompt, makes explicit the latent criteria a high-level intent entails (DG1) and lets a moderator attribute a given verdict to a specific criterion rather than to an opaque model (DG2). It also lets a moderator trace a verdict back to the criterion and the rule text that produced it, supporting transparent and traceable decisions (DG4).

The two forms are kept aligned in one direction: the automoderator logic is compiled from the rule text, which serves as the source of truth. When a moderator edits the rule text, DecideTogether recompiles the affected logic so that the executable criteria continue to reflect the stated rule. This directionality lets a moderator revise rules in natural language and follow how each revision alters concrete decisions, without operating on the logic directly (DG3). A moderator can still edit the logic directly when a criterion needs an adjustment that the rule text does not capture, but the rule text remains the artifact the alignment is anchored to.

6.3.2 User Interaction

We describe how a moderator uses DecideTogether to turn an initial, often vague moderation intent into a rule whose enforcement consequences they have examined and endorsed. The interaction is not a one-shot specification but a loop: the moderator establishes a shared account of the community, articulates an intent into a rule text and its compiled logic, tests that logic against real cases, and revises. We organize this section around the four stages of that loop and describe how each stage supports the moderator in articulating intent.

Establishing Shared Context

Upon entering the system, the moderator first sees the Community Profile page (Figure 6.1). For each community, DecideTogether samples representative posts and comments and infers a profile across four dimensions: Purpose, Participants, Stakes, and Tone (Section 6.3.3), presenting each dimension as a set of editable tags. The moderator reviews this analysis and corrects any tag that does not align with their own understanding of the community.

This stage establishes the community’s shared account, on which every subsequent rule is based. A moderation intent is rarely community-neutral, and the same behavior may warrant different treatment depending on who participates and what is at stake. Surfacing and editing this context before any rule is written gives later compilation a shared, inspectable account of the community to draw on, which is the community-level half of surfacing the latent criteria an intent entails (DG1).

Articulating Intent into Rule Text and Logic

To add a rule, the moderator opens the New Rule modal (Figure 6.2) and describes their intent as a rule title, then selects the community contexts relevant to the rule with the sliders. To help the moderator start from concrete precedent rather than a blank slate, DecideTogether retrieves rules from communities that share context clusters and presents them as references labeled by source community and shared context tags. Because a vague intent is hard to translate into enforceable criteria unaided, DecideTogether then generates a draft of the rule text from the title and selected contexts, which the moderator can accept or revise.

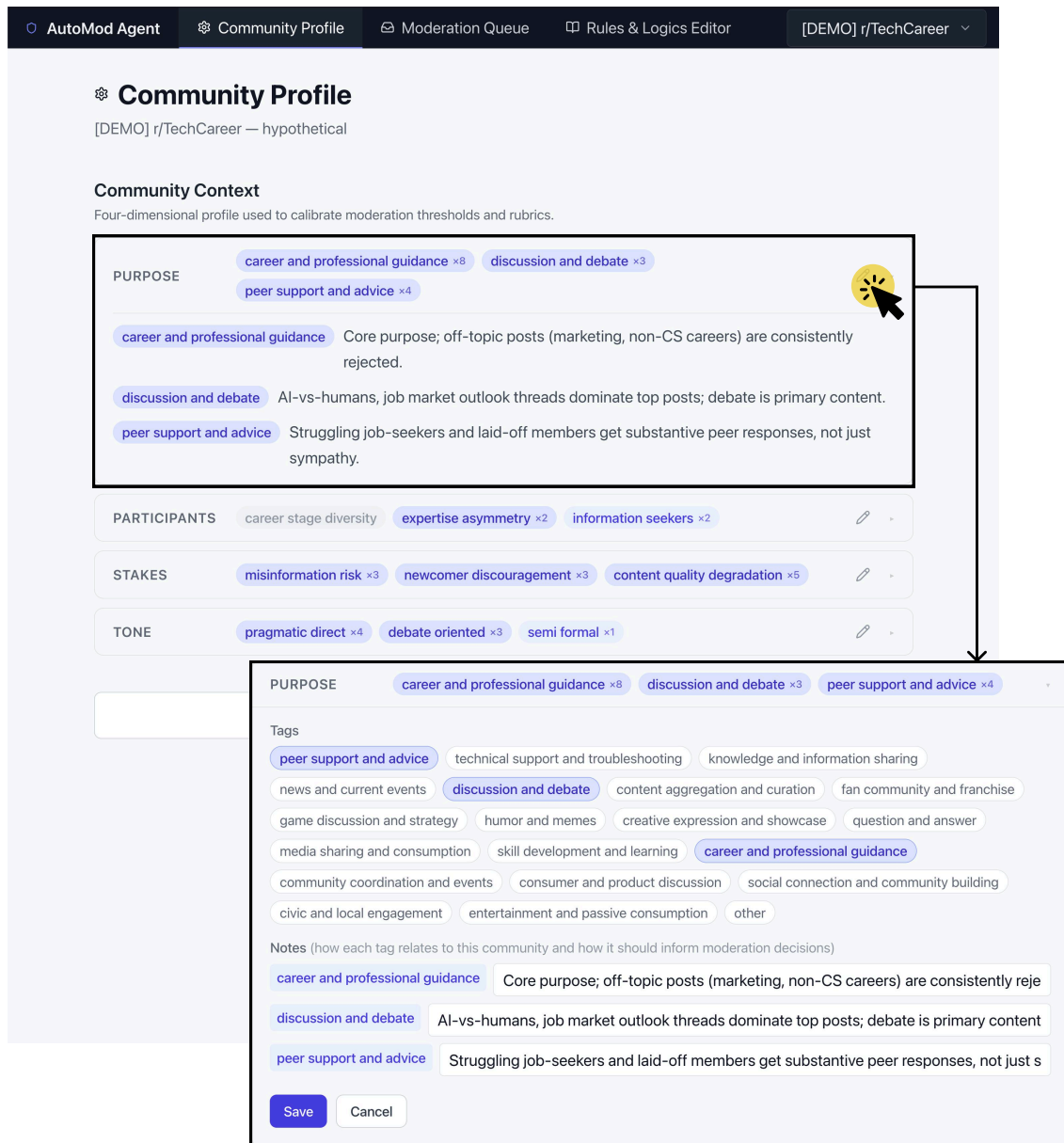


Figure 6.1: In the community profile page, DecideTogether shows the characteristics of the community as tags across four dimensions: Purpose, Participants, Stakes, and Tone. The user can further edit the community profile.

New Rule

A

Title

Comments must be professional and helpful

B

Relevant context tags

Pick to scope suggestions — 4 selected

PURPOSE

career and professional guidance

strongly informs

peer support and advice

strongly informs

discussion and debate

ignore

PARTICIPANTS

career stage diversity

ignore

information seekers

ignore

expertise asymmetry

ignore

STAKES

misinformation risk

ignore

content quality degradation

ignore

newcomer discouragement

informs

TONE

pragmatic direct

informs

semi formal

ignore

debate oriented

ignore

C

Compare options · LLM draft + 0 peer rules

Regenerate

No closely-matching rules from peer communities — this rule looks niche. The drafted option below is grounded only in your community context.

Each option shows a rule and the community context it grew out of. Page through to compare framings; pick the one whose context resonates, or refine from there.

Drafted from your context

Use this

- All comments must be directly relevant to tech careers, job-seeking, or professional development in CS-related fields — off-topic responses (e.g. marketing pitches, advice about unrelated industries) will be removed.
- Responses to members seeking career guidance must include substantive, actionable advice; replies that offer only sympathy or emotional validation without practical content are not sufficient.
- While blunt and direct feedback is welcome and expected, comments must not veer into unnecessarily discouraging or doom-framing language that could distort the career outlook of newer members.
- Personal attacks, demeaning language, or any content that undermines a professional and constructive discussion environment will result in comment removal and may lead to a ban.

career and professional guidance

peer support and advice

newcomer discouragement

pragmatic direct

Synthesised from your community context + 0 peer rule(s)

Figure 6.2: New Rule page of DecideTogether. (A) The user can elaborate on their initial moderation intent as the rule title. (B) The user can select the relevant community profile tags. (C) Based on the user’s moderation intent and community contexts, DecideTogether generates a draft of the rule text.

94

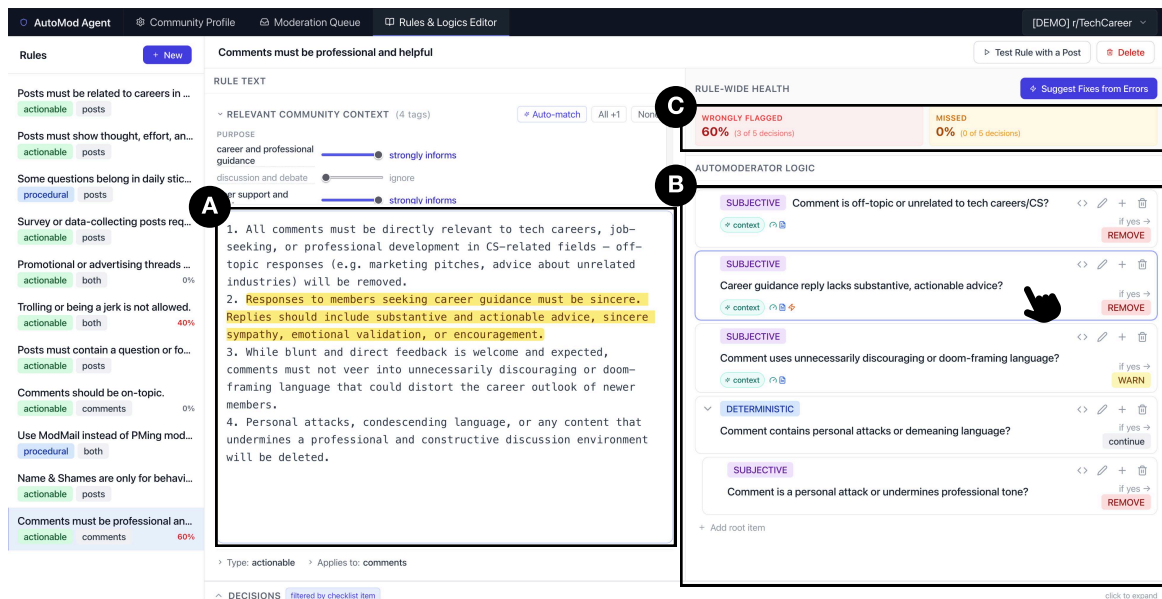


Figure 6.3: Rule editor page of DecideTogether. (A) DecideTogether shows the rule text on the left. (B) On the right, DecideTogether shows the automoderator logic constructed from the rule text. (C) Hovering over each logic item highlights the specific part of the text it is relevant to. Furthermore, it shows the number of incorrectly flagged and missed instances to help users assess the accuracy of the automoderator.

On creation, DecideTogether compiles the rule into the set of automoderator logic items described in Section 6.3.1, shown alongside the rule text in the rule editor (Figure 6.3). Hovering on a logic item highlights the part of the rule text it came from, exposing the connection between the two forms (Figure 6.3B).

This pairing is where a tacit intent is first made explicit at the level of the rule itself. The compiled logic surfaces the criteria the rule entails, complementing the community-level context from the previous stage (DG1). By tying each logic item back to the rule text, the stage lets the moderator read the logic not as a black box but as an externalized account of their own intent that they can check.

Testing Intent Against Real Cases

The moderator evaluates the system’s decisions and issues final verdicts on the moderation queue page (Figure 6.4). Each item shows the post or comment, the system’s decision, and the reasoning produced by each logic item, so the moderator can trace why DecideTogether reached a given verdict (Figure 6.4B). The queue also surfaces posts that break no explicit rules but poorly fit the community’s norms, so the moderator can catch problems the current rules do not yet cover. The moderator issues a final verdict of approve, warn, or reject; when overriding the system, they can leave a short note explaining the decision and link it to the specific rule violated (Figure 6.4A).

This stage turns the rule into evidence about whether it captures the intent. Because each verdict can be traced back to the criterion and the rule text that produced it, a mismatch can be attributed to a specific criterion rather than to an opaque model, which makes the result usable as a signal for revision rather than an outcome to be debugged (DG2). The same per-item reasoning is what makes a

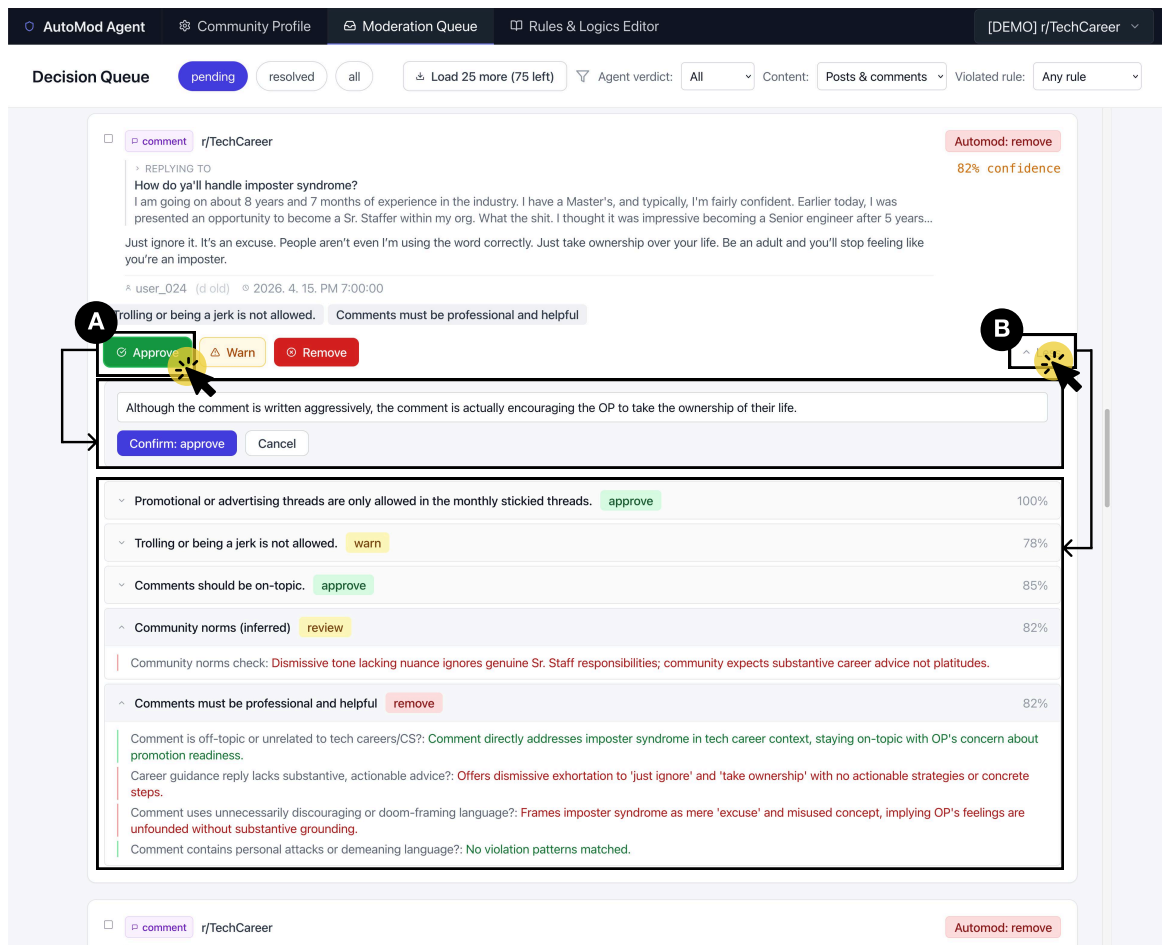


Figure 6.4: Moderation queue of DecideTogether. (A) When the user is overriding DecideTogether's decision, the system elicits the reasoning. (B) For each automoderator logic, DecideTogether shows the reasoning behind the decision.

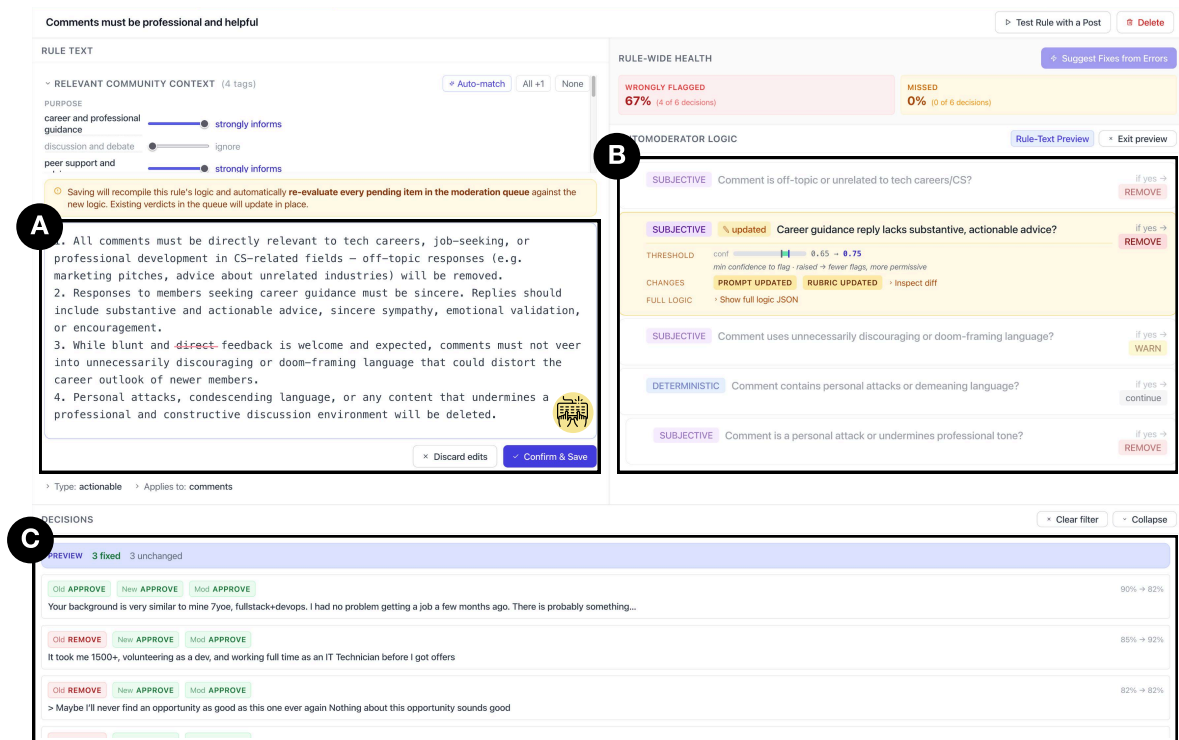


Figure 6.5: Interaction for editing the rule text on DecideTogether. (A) When the user edits the rule text, DecideTogether automatically shows (B) a preview on the logic update and (C) how the new logic would affect the previous automoderator decisions.

contestable decision defensible, since the moderator can show the criterion and the rule text that support it (DG4). The override notes left here are retained as the signal that drives the revision stage.

Revising Intent

To align the automoderator’s decisions with their own criteria, the moderator reviews them in accessible terms rather than operating on the logic directly. First, the moderator can edit the rule text to clarify their intent (Figure 6.5). DecideTogether recompiles the rule, previews the resulting changes to the logic (Figure 6.5B), and re-evaluates previously decided cases to show how the verdicts would change under the new logic (Figure 6.5C). The moderator can also adjust the community contexts relevant to the rule, upon which DecideTogether suggests updates to the rule text and recompiles the logic accordingly.

Second, the moderator can click “Suggest Fixes from Errors,” which draws on the wrongly flagged cases, missed cases, and override notes gathered during testing to infer updates to the logic (Figure 6.6). DecideTogether previews each fix on the rule text (Figure 6.6B), the logic (Figure 6.6C), and the previously decided cases (Figure 6.6D), so the moderator can judge it before accepting. If a criterion needs an adjustment that the rule text does not capture, the moderator can edit the logic items directly, including the rubrics, LLM prompts, or regular expressions underneath them, though the rule text remains the artifact to which the alignment is anchored.

By making every revision proceed through accessible edits whose effects on concrete decisions are shown, this stage keeps the consequences of each change legible without requiring the moderator to author logic directly (DG3).

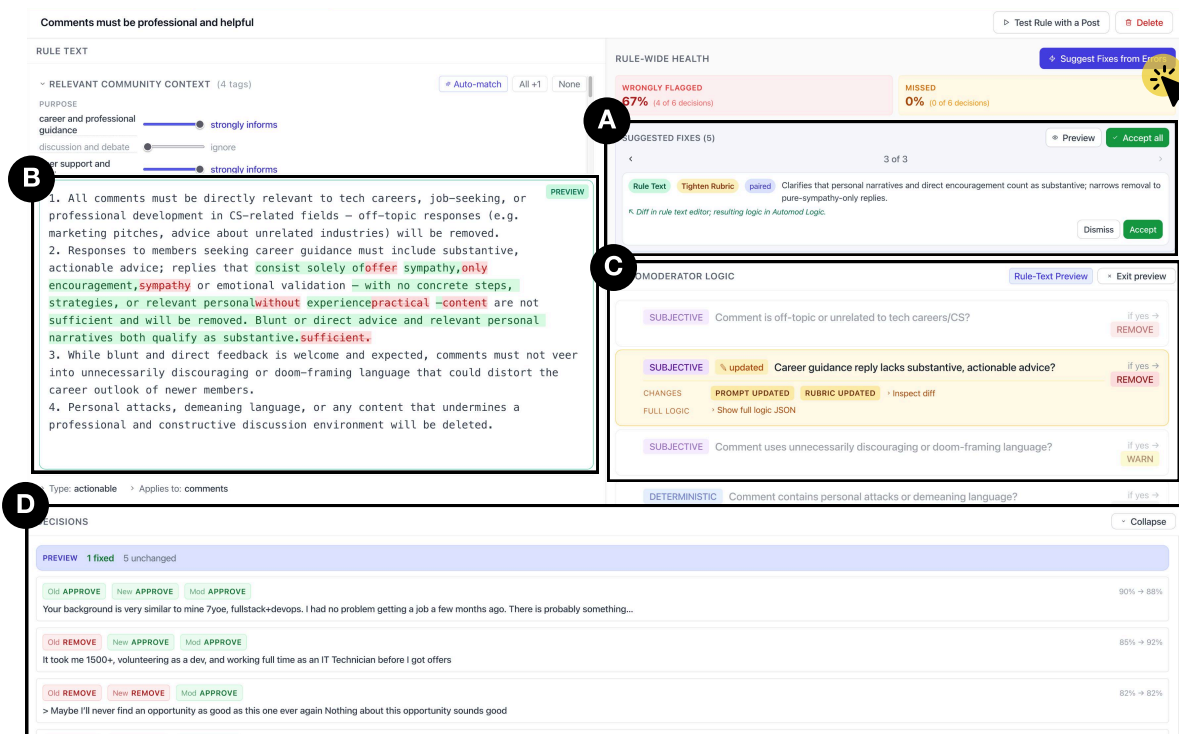


Figure 6.6: Interaction for getting suggestions to update the automoderator logic on DecideTogether. (A) DecideTogether shows a summary of each rule update. DecideTogether further shows (B) the effect of the fix on the rule text, (C) automoderator logic, and (D) the previous automoderator decisions.

6.3.3 Technical Pipeline

DecideTogether implements several LLM-powered pipelines to power the interactions. We describe the details of each pipeline component.

Community Context Inference

DecideTogether takes a set of sample posts and comments to infer the community contexts. First, to define a set of community characteristics, we defined the four dimensions: Purpose, Participants, Stakes, and Tone. Purpose describes the goals and the purposes of the people participating in the community. Participant describes the characteristics of the community members. Stakes describes the potential risks the community may face and the potential harm it could cause if moderation fails. Tone describes the characteristics of interaction styles and languages.

To ground the structured tag vocabulary in real community variation rather than designer intuition, we derived the taxonomy empirically from a corpus of 1,000 subreddits sampled via Reddit’s public API, combining `/subreddits/popular` (the 1,000 highest-traffic communities) with `/subreddits/search` across 72 topical queries (e.g., mental health, legal advice, woodworking, gaming) to ensure breadth across purpose and participant types. For each subreddit, we prompted Claude Haiku with the sidebar description and rule text in an open-ended extraction task, asking it to produce a short prose summary and 3–5 free-form tags along four dimensions, Purpose, Participants, Stakes, and Tone, without any predefined categories, yielding roughly 2,000–2,700 unique tags per dimension. We then canonicalized this long tail through a two-pass clustering procedure with Claude Sonnet: a discovery pass over the 150 most frequent tags per dimension produced 10–25 candidate categories with descriptions, and a mapping pass assigned all unique tags to canonical categories in batches of 200, collapsing near-duplicates and absorbing rare tags into meaningful groups. We validated the resulting taxonomy along five axes: residual “Other” bucket size ($\leq 5\%$ on every dimension), co-occurrence patterns, category balance, saturation (no new categories emerged after 100–200 communities), and qualitative spot checks (e.g., `r/wallstreet-bets`: `humor_heavy`, `dark_or_edgy_humor`, `financial_harm_risk`; `r/mentalhealth`: `supportive_encouraging`, `vulnerable_populations`). The final taxonomy contains 20 Purpose, 23 Participants, 23 Stakes, and 19 Tone categories.

When analyzing the community profile, DecideTogether samples up to 65 posts and 50 comments. A community’s norms are visible not only in what it rewards but also in what it ignores, so DecideTogether samples across the full range of engagement rather than only its most visible content. It draws up to 25 hot posts for typical day-to-day content, up to 10 top, and up to 10 controversial posts from the past month for the upper and contested ends of reception, and up to 20 ignored posts that drew little engagement (an upvote ratio below 0.45 and below-median score and comment counts) as negative examples. For comments, it takes the top five comments from each of the first ten hot posts. Based on the samples, DecideTogether uses Claude Sonnet 4.6 to assign 2–3 tags from each dimension and generate a short note on how each tag relates to the characteristics of the community. This inferred profile is the community-level context against which a rule’s criteria are later surfaced (DG1).

Compiling Community Rules

The automoderator logic introduced in Section 6.3.1 consists of a set of logic items, each comprising a binary question, an internal logic, and an action to be taken if the question is answered affirmatively. The objective criteria are implemented as two item types, and the subjective criteria as a third. Deterministic

items handle clear-cut keywords and pattern-level filtering, using a set of regular expressions to detect violating patterns in posts or comments. Structural items use platform-specific signals, such as the number of upvotes or account information, to filter out spam or bot accounts. Subjective items filter posts on non-objective criteria, such as intent or message tone; internally, such an item includes an LLM prompt and a rubric, and is evaluated by running the prompt against the rubric. To balance cost and accuracy, DecideTogether evaluates subjective items in two tiers: it first evaluates all subjective items for a post in a batch with Claude Haiku 4.5, then re-evaluates any item whose confidence falls below a threshold (0.75) with Claude Sonnet 4.6.

To accommodate the contexts, DecideTogether uses a two-pass compile approach with Claude Sonnet 4.6. In the first pass, DecideTogether uses an LLM prompt that takes only the rule title and content as input to produce the initial set of logic items. In the second pass, DecideTogether runs a second LLM prompt to tune the initial set based on the user-provided community contexts, which may add a new item, remove one, or update the internals of a logic item. Decomposing the rule into discrete items, rather than a single prompt, allows a verdict to be attributed to a specific criterion and traced back to the rule text (DG2, DG4).

Flagging Context Misfit

Not every problem a community cares about is covered by an explicit rule. A post may break no stated rule yet still sit poorly with the community’s norms, given who participates and what is at stake. To moderate for this, DecideTogether runs a separate evaluation path for posts that violate no explicit rule, using Claude Haiku 4.5 to judge whether a post is a poor fit for the community profile and to flag it for the moderator’s review. This does not bypass the rule text as the source of truth: a flag is not a verdict but a candidate the rules do not yet cover, surfaced for the moderator to judge and, if warranted, turn into a rule. This path applies the same inferred community context that grounds rule compilation, extending that context beyond the rules a moderator has written so far (DG1).

Generating Logic Update Suggestions

To support users in updating the logic to better align with their own decisions without manually editing the rule text and the logic, DecideTogether analyzes wrongly flagged and missed cases to generate the updates. DecideTogether uses Claude Sonnet 4.6 to diagnose the mismatches between the current logic and the user-made decisions and suggests updates that preserve the structure of the original auto-moderator logic. The prompt takes the rule text, the current status of the logic, community contexts, sibling rules, the actual content of the wrongly flagged and missed cases, and the user-attached notes for those cases. The prompt gives suggestions at several different levels. First, the prompt proposes updates to the logic within the following action space:

- Tighten rubric: When the current logic gives inconsistent results, DecideTogether suggests clarifying the logic.
- Adjust threshold: When the current logic is too sensitive or lenient, DecideTogether suggests tuning the threshold.
- Split item: When the current logic conflates two distinct concepts and gives inconsistent results, DecideTogether suggests splitting the logic into distinct items.

- Add item: When there are user-annotated violations that the current logic does not capture, DecideTogether suggests adding a new logic item.

For each logic update, DecideTogether additionally proposes an update to the rule text to keep the rule text, which remains the source of truth, consistent with the accepted logic change. By proposing these updates as accessible edits whose effect is previewed before acceptance, this component lets the moderator revise without authoring logic directly (DG3).

Rule Text Suggestion

To help a moderator draft a rule grounded in real moderation practice rather than from scratch, DecideTogether suggests rule text by drawing on how comparable communities have written similar rules. We built a reference corpus of 100 communities, sampled from the AIRules dataset [240] and stratified across five community archetypes. We specifically selected communities with more than 10,000 subscribers and substantive rulesets consisting of 5-15 rules, each with descriptions of more than 50 characters. For each community, we applied the same context-inference pipeline to tag it with a four-dimensional community profile and to embed each rule title. When a moderator enters a new rule title, DecideTogether embeds the title and retrieves reference rules with the highest cosine similarity, keeping up to 5 peer rules from distinct communities that exceed the similarity threshold. DecideTogether then passes these peer rules together with the target community’s context to Claude Sonnet 4.6, which drafts a rule in which every clause must cite at least one grounding source, either a tag from the target community context or one of the peer rules. When no reference rule clears the similarity threshold, DecideTogether drafts from the target community’s context alone, grounding every clause in its context tags. DecideTogether returns both the drafted rule and the peer rules themselves, so the moderator can start from a concrete draft or browse the peer rules as starting points.

6.4 Technical Evaluation

To evaluate whether our technical pipeline accurately captured moderation intents and made accurate moderation decisions, we conducted a series of technical evaluations. We specifically evaluated the following questions:

- Q1. Does automoderation logic generated by DecideTogether accurately capture the moderation intent of the rule text?
- Q2. Do the moderation decisions from the automoderator logic generated by DecideTogether match with the decisions of the human moderators?
- Q3. Does updating the automoderation logic generated by DecideTogether lead to better alignment with human decisions?

To evaluate the moderation decisions of DecideTogether, we leveraged the existing dataset of moderated comments [222]. The dataset contains around 2M removed comments from the top 100 subreddits from May 2016 to March 2017. Among the subreddits, we selected 15 text-heavy subreddits across diverse topics. The details of the selected subreddits are shown in Table 6.1.

To construct the dataset for each subreddit, we first randomly sampled 50 removed comments from the moderated comments dataset, excluding comments with body text of fewer than 20 characters or

Table 6.1: The set of subreddits we used for the technical evaluation, and the number of rules for each subreddit

Subreddit	Number of Rules
hiphopheads	34
AskHistorians	30
relationships	25
legaladvice	22
space	20
anime	14
science	12
personalfinance	10
changemyview	10
news	10
AskReddit	9
explainlikeimfive	7
asoiaf	7
books	6
depression	4
Total	220

containing the text `[deleted]` or `[removed]`. To augment the rule-compliant comments, we sampled around 100 comments from the PushShift Reddit archive for the same period, filtering short/deleted bodies, automoderator bot comments, and `[deleted]` authors. We also re-validated whether each comment was not moderated after the PushShift crawl and removed the moderated ones. In total, the ground-truth dataset contained 2,146 comments across 15 subreddits.

To reconstruct the ruleset for each subreddit, we retrieved each subreddit’s ruleset from the Internet Archive [241] in March 2017. Reddit had its “old” interface in March 2017, each community using different designs and structures to share their rulesets. We manually parsed the ruleset to organize it into rule titles and rule texts. Table 6.1 shows the number of rules for each subreddit. The number of rules and their granularity significantly differed: `r/depression` had only 4 rules in 2017, while `r/hiphopheads` had 34 rules. We also reconstructed the community context during the same period by sampling posts and comments from that period.

To answer Q1, we used an LLM-as-a-judge approach to evaluate whether the generated automoderator logic accurately captured the moderation intent embedded in the rule text. We specifically used the following four measures:

- Logical correctness: Does the automoderator logic faithfully represent the original rule without contradictions or misinterpretations?
- Clarity: Is each of the questions phrased in a way that a human moderator could answer them without ambiguity?
- Coverage: Does the automoderator logic cover all distinct requirements in the rule?
- Minimality: Is the automoderator logic free of unnecessary or duplicate conditions?

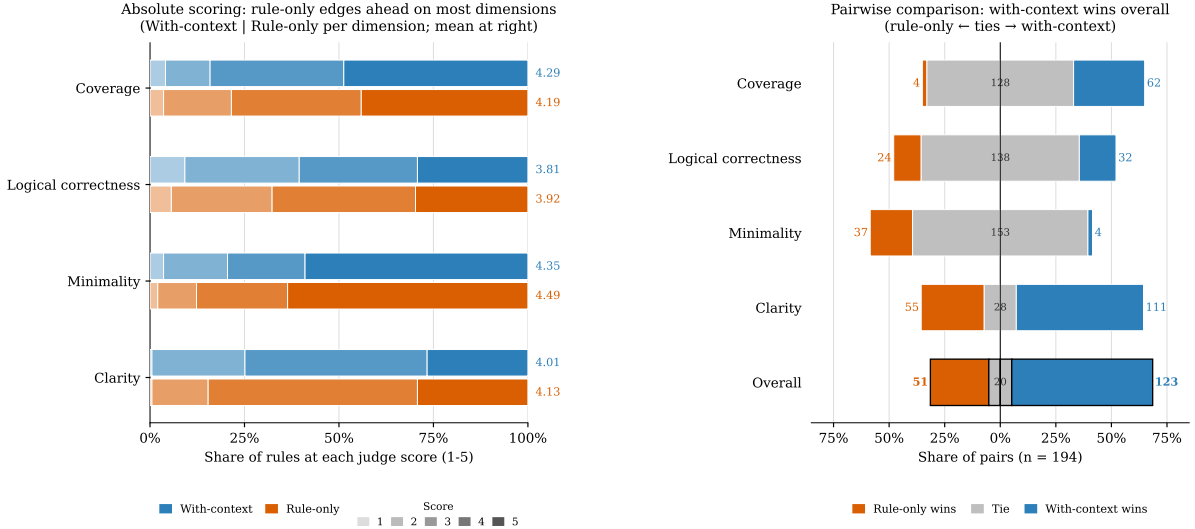


Figure 6.7: Comparison of the quality of automoderator logic generated with and without contexts. Absolute scores from With-context and Rule-only conditions were comparable, but With-context conditions won in pairwise comparisons.

For each measure, we prompted a **Claude-Sonnet-4.6** model to rate the quality of the automoderator logic on a 5-point Likert scale. We generated the logic with and without the community context information and compared their scores. Additionally, we used the same LLM model to compare the quality of automoderator logic pairwise.

To answer Q2, we computed the accuracy, precision, recall, and F1 score on the ground-truth dataset of moderated comments. We compared the performance metrics among three conditions: Direct, Rule-Only, and With-Context. In the Direct condition, we directly fed the whole ruleset as the prompt to the LLM. In the Rule-only Logic condition, we used the DecideTogether-generated logic without the community contexts. In the With-context Logic condition, we used DecideTogether-generated logic with the community context information.

To answer Q3, we created another condition: With-Update. For the With-Update condition, we ran the automoderator logic from the With-Context condition on a second set of ground-truth comments of approximately the same size, constructed with the same sampling strategy. Based on the results, we generated and applied fixes to the automoderator logics and ran the evaluation again on the same set.

6.4.1 Q1: Quality of Automoderator Logic

The results suggest that the pipeline for DecideTogether produced automoderator logic that aligned with the moderation intents embedded in the rule text, and adding community contexts led to better alignment with those intents (Figure 6.7). In terms of absolute scoring, rule-only logic scores marginally higher on logical correctness (3.92 vs. 3.81) and clarity (4.13 vs. 4.01). In paired comparison, however, with-context wins overall 63% of the time (vs. 26% for rule-only, 10% ties; $n=194$), with large wins on clarity (57% vs. 28%) and coverage (32% vs. 2%). Minimality is the only dimension where rule-only wins head-to-head (37 vs 4). We expect that context-driven compilation would lead to adding more automoderator logic to the rule, potentially introducing unnecessary complexity.

Table 6.2: Functional accuracy of the DecideTogether-generated automoderation logic. (15 subreddits, $n=2146$). Best non-trivial value per column in **bold**.

Condition	Acc.	κ	Remove class	Accuracy by source	
			P / R / F1	removal_log ($n=750$)	pushshift ($n=1396$)
<i>Always approve</i>	<i>65.1</i>	<i>.00</i>	<i>- / .00 / -</i>	-	-
<i>Always remove</i>	<i>35.0</i>	<i>.00</i>	<i>.35 / 1.00 / .52</i>	-	-
Direct (no logic)	45.9	.084	.38 / .87 / .53	87.5	23.5
Rule-only logic	65.4	.29	.50 / .67 / .57	66.7	64.7
With-context logic	64.5	.27	.49 / .63 / .55	62.8	65.5
With-update logic	65.9	.32	.54 / .62 / .58	62.3	71.0

κ interpretation: .09 “slight”, .27–.32 “fair”.

6.4.2 Q2 and Q3: Moderation Performance

Table 6.2 shows the performance of the automoderation logic. The results show that DecideTogether’s approach of decomposing rule text into automoderator logic substantially improved the accuracy compared to direct rule-text prompting. While the aggregate accuracy is 65.4% for Rule-Only condition and 64.5% for With-Context condition, Direct condition had only 45.9% of accuracy. Because the ground-truth class distribution is skewed (65% approve / 35% remove), we also report Cohen’s κ to control for chance agreement. While Rule-Only logic reaches $\kappa = 0.29$ and With-Context logic reaches $\kappa = 0.27$, direct reaches $\kappa = 0.084$, suggesting that decomposing into automoderator logics led to better agreement with the ground truth. Comparison of precision and recall shows that the Direct approach was excessively deciding to remove comments, leading to higher recall with the removal decisions but poor performance with accepting comments.

At the same time, direct comparison between Rule-Only and With-Context conditions shows that the Rule-Only condition performed slightly better. Per-subreddit analysis shows a more fine-grained picture, suggesting the community norm and the granularity of existing rules affected the effect of contexts. Figure 6.8 shows the difference in accuracy before and after adding the community contexts per subreddit. For subreddits like r/science, r/depression, and r/space, adding community contexts actually improved the moderation accuracy. Such communities often had underspecified rules. For instance, each rule in r/science was only in one line. In such cases, adding community context actually helped calibrate decisions by editing the rubrics to restore falsely removed comments, clarify the scopes for each rule, and establish clearer moderation criteria. On the other hand, for subreddits like r/AskHistorians, r/legaladvice, and r/personalfinances, adding context led to a decrease in performance. Such communities often had stricter community norms on the contribution quality and potential risks as well as stricter rules. In that case, adding community context tightened the rubric and added more criteria to filter out comments, resulting in fewer comments being removed.

Furthermore, the accuracy numbers suggested that DecideTogether updating the automoderator logic based on errors improved automoderation performance in general. Figure 6.9 shows the difference in accuracy before and after updating the automoderator logic. For most subreddits, updating the logic generally increased accuracy. However, for some cases, like r/AskHistorian, updating the logic led to worse performance. In the case of r/AskHistorian, DecideTogether already implemented complex

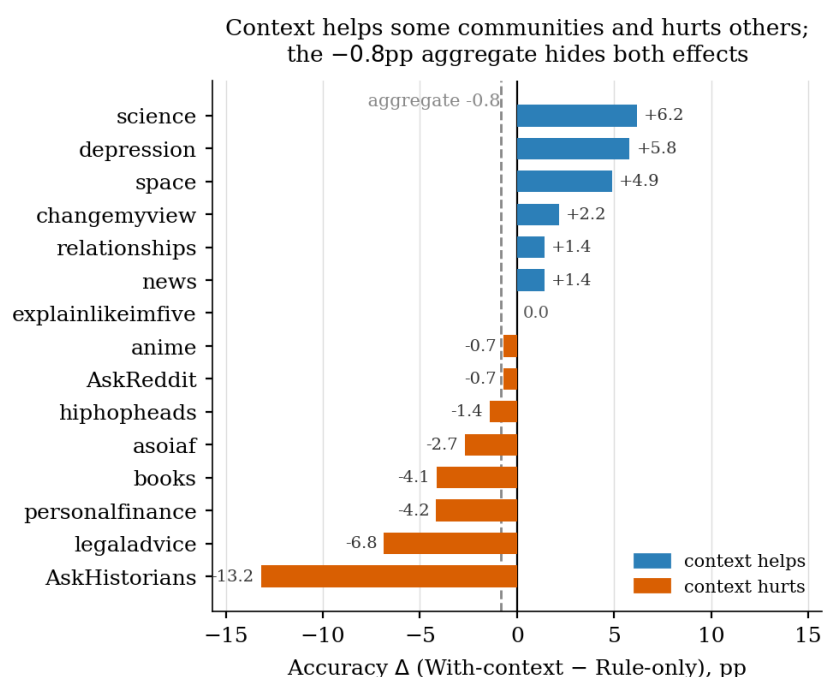


Figure 6.8: Per-subreddit difference in overall accuracy before and after adding the community context to the automoderator logic. While adding context improved accuracy for communities with underspecified rules, such as r/science and r/depression, it reduced accuracy for communities with stricter norms, such as r/AskHistorians and r/legaladvice.

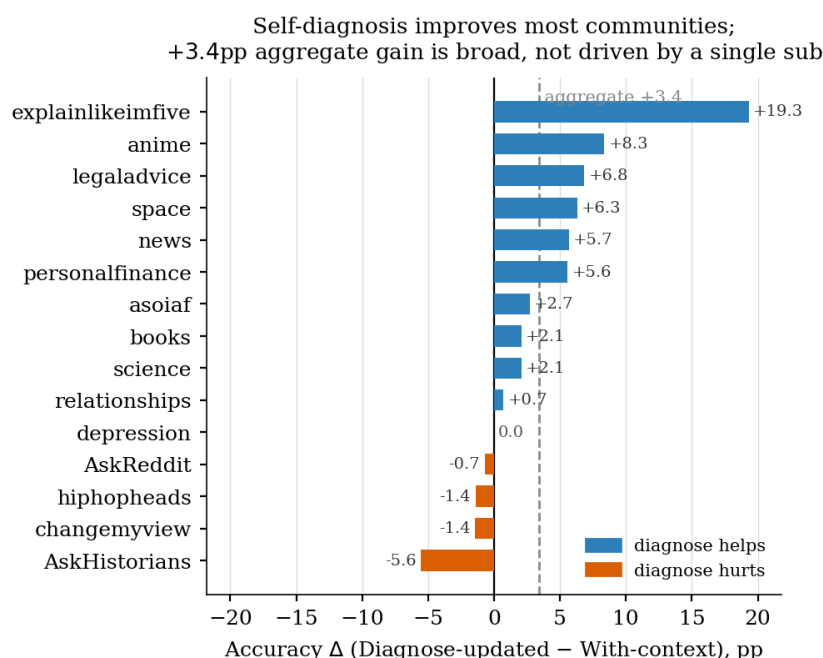


Figure 6.9: Per-subreddit difference in overall accuracy before and after updating the automoderator logic from its errors. While updating the logic increased accuracy for most subreddits, it reduced accuracy for communities with already complex logic, such as r/AskHistorians.

Table 6.3: Previous moderation experiences of the user study participants

Participant No.	Main platform	Monthly active users	Prior moderation experience	Scenario
P1	Discord	500	3-5 years	HotTakes
P2	Discord	600	1-3 years	AskAScientist
P3	Discord	3,000	5 years +	TechCareers
P4	Twitter groups	472,000	5 years	HotTakes
P5	Reddit	9,400	1-3 years	HotTakes
P6	Facebook groups	26,000	5 years	AskAScientist
P7	Reddit	1 million +	5 years	AskAScientist
P8	Open group chat	35	1-3 years	TechCareers
P9	Discord	2,500	1-3 years	TechCareers
P10	Discord	250,000	5 years	HotTakes

automoderator logic, and adding the updates further complicated it.

6.5 User Study

To understand how DecideTogether could support expressing and refining the moderator’s moderation intents, we conducted a user study. We specifically investigate the following research questions.

RQ1 How does DecideTogether support community moderators in articulating and refining moderation intent?

RQ2 How does DecideTogether affect moderators’ understanding, trust, and control over automoderation rules?

6.5.1 Participants

We recruited 10 participants from cold-mailing Reddit moderators via modmail, sharing open call on the researcher’s social media, and from Upwork. We specifically aimed to recruit participants with previous community moderation experiences, regardless of platform. Table 6.3 shows the details of the participants’ information.

6.5.2 Scenarios

For the user study, we developed three scenarios, each describing a problem with a hypothetical community. We designed the scenarios to be comprehensive in covering diverse issues reported in the existing literature on automoderators [242, 226, 226] as well as diverse types of communities, including Q&A, discussion, and advice-seeking. Each hypothetical community was based on an actual, active subreddit.

1. Scenario 1. r/HotTakes is a casual community for sharing genuine minority opinions that others can engage with by agreeing or pushing back. Posts frequently disguise questions, rants, observations, or advice as opinions, leaving no clear position for the community to engage with or disagree against. (Base community: r/unpopularopinions)

2. Scenario 2. `r/AskAScientist` is a community for asking scientific questions and receiving evidence-based answers from verified experts and credentialed researchers. Speculative and hypothetical questions that cannot be answered by scientific literature are crowding out genuine, answerable science questions and inviting non-expert speculation. (Base community: `r/askscience`)
3. Scenario 3. `r/TechCareers` is a community for discussing software engineering and CS careers — job searching, salary, workplace issues — with an emphasis on practical, experience-based advice. The community values frank, direct advice, but comments increasingly cross into dismissiveness and gatekeeping, discouraging newer members without clear rules distinguishing candor from cruelty. (Base community: `r/cscareerquestions`)

Each hypothetical community consisted of a ruleset, a set of posts or comments to be moderated, and another set of sample posts and comments for community profile construction, to show the current state of the community, including problematic posts or comments. The ruleset was adopted from the base community’s ruleset, excluding any rules that addressed the problematic cases in the scenario. The set of posts or comments requiring moderation included both scenario-compliant and scenario-violating contributions. To populate the scenario-compliant examples, we first crawled live posts and comments from the base community and filtered them using the existing ruleset to ensure contributions did not violate rules irrelevant to the scenario. Scenario-violating examples were challenging to crawl, as they were often removed by moderation before we could crawl them. Therefore, we decided to generate synthetic examples using the Claude-Opus-4.7 model. The prompt used the scenario, existing ruleset, and the taxonomy of violations as the input. We also added a small set of human-curated examples of scenario-compliant, scenario-violating, and borderline cases, extracted from the ArcticShift dataset. We instructed the model to generate examples that match the community tone, length, and vocabulary while not being too obvious violations. We sampled 75 scenario-compliant examples and 25 scenario-violating examples to construct a final dataset of 100 contributions. The set of sample posts and comments for community profile construction was directly crawled from the base community.

6.5.3 Procedure

Before starting the task, we provided a step-by-step introduction to the system to explain its workflow. After that, we invited the participants to try out the system using a demo scenario to understand how it works.

Then, we introduced the task scenario and the main task to the participant. In the task, we specifically gave a four-step workflow. First, we asked participants to skim sample posts and comments to understand the community’s norms and culture and to verify the community profile. Then, based on their understanding of the community context and the scenario, participants drafted the rule text with their initial moderation goal in mind using DecideTogether. After DecideTogether produced moderation decisions on the set of posts and comments, they were invited to review them and identify any cases where DecideTogether’s decision did not match with their own. Then, the participants were asked to tweak the rule texts and the automoderation logic so that DecideTogether’s moderation decisions would align with the participants’ own moderation goals. The task lasted 40 minutes, and participants were asked to complete the first part within the first 10 minutes.

After the task, the participants responded to a semi-structured interview to share their experience with DecideTogether in articulating their moderation goals into rules and in calibrating their trust in the system.

6.6 Results

Participants followed almost the same user flow as instructed by the task. They started by writing a short natural-language moderation intent and accepted the system-generated rule text. Then, from the moderation queue, they discovered misalignments and missed cases of the automoderator, which led to refining the rule in response. Here, we report the findings from our user evaluation.

6.6.1 RQ1: How does DecideTogether support community moderators in articulating and refining moderation intent?

We frame moderators’ engagement with DecideTogether as a bidirectional dialogue in which different system representations support them in articulating, testing, and refining intent that was initially vague or tacit. We explain the participants’ usage patterns in three stages: Drafting the initial rule, Reflecting on the moderation intent, and Revising the moderation intent.

Establishing Shared Understanding Before any rule was authored, the system-generated community profile established the shared contexts between DecideTogether and the moderators about the community itself, against which subsequent rules would be drafted. This shared grounding operated in two recurring ways. First, it served as a verification surface: moderators read the profile, judged it against the sample posts, and made targeted edits, explicitly recognizing this as a communicative act (P3: “I’m writing this so the LLM has more information about the community... giving context to LLM”). Edits are often propagated directly into subsequent rule decisions (P2). Second, it served as a direct input channel: by adjusting sliders and weighting tags, moderators conveyed their priorities around the rule (P9: “when I get the slider in which area I can strongly inform or ignore, overall that makes it much better to my comment to the automated bot”). Notably, this channel was concentrated at the drafting stage; once the rule text existed, moderators rarely returned to it.

The profile’s role as shared context, however, was not frictionless. Three limitations recurred. First, the profile felt too abstract at first encounter, particularly for moderators unsure how to engage with tag-based representations (P6: “a bit too like blank”; P7: “too calculative... make it less abstract, more comprehensible”). Second, the profile became invisible during rule authoring — once moderators left the profile page, the content continued to shape LLM output but was no longer co-present to them (P8: “I was probably influenced unconsciously, but I didn’t consciously use it... it would be nice to have small reminders next to the rule generator”). Third, several moderators recognized the inherent incompleteness of automated profiling: community character cannot be fully recovered from sampled posts (P9: “Automated moderation or bot cannot be 100% accurate... some pointers may be missed and maybe still reviewed by a human moderator”). Together, these point toward keeping the shared context visibly co-present and humanly verifiable throughout the dialogue, not only at its opening.

Drafting DecideTogether supported participants to express their moderation intents in rule titles within relevant community context, which were elaborated into fully written rule texts generated by DecideTogether. Participants valued the DecideTogether-generated rule texts for accurately capturing and concretizing their moderation intent within the community contexts. P8 said, “Because the system and I already shared the same community context, even with a simple prompt, the system understood the characteristics of the community’s ecosystem well, and that made me feel like it really understood what I meant.” P5 and P9 considered that DecideTogether-generated rule texts surfaced ideas and

considerations that they couldn't articulate by themselves. P9 said "it skyrocketed my expectation... even the part of context which I cannot think of, it automatically added itself in a descriptive way". With such a high trust on the DecideTogether-generated rule texts, most participants decided to accept the DecideTogether-generated rule texts directly with minimum modification. Participants also commonly valued DecideTogether that they could easily express their moderation intents, especially as they do not have a technical background to set up complex rules for automoderators.

Reflecting From the rule text, DecideTogether interpreted the moderation intent into automoderator logics and showed how they would be enacted with actual posts and comments.

DecideTogether-generated automoderator logics enabled the participants to reflect on what their rule actually meant. P7 described the experience with the automoderator logics, saying "it almost explained it back to me in a way." The decisions on the moderation queue supported the participants to reflect on what their rule actually does in the community contexts. DecideTogether often produced hard-and-fast automoderator logic that strictly adhered to the rule text. After reviewing DecideTogether's decisions, participants learned that the rule should be more lenient to accommodate contributions that would genuinely benefit the community. P6 shared their task experience with making a rule on "Every statement should be backed up with real data" to moderate hypothetical questions, saying "I changed the logic to be a little bit more lenient around studies. Some posts were harmless, but because they didn't mention a study, the automod flagged them as potential misinformation. But actually, I know as a human that the person was asking for information and real studies." P3 specifically explained that the reasoning for each of the automoderator's decisions was helpful, "When you are moderating conversations, sometimes you get the feeling that something is wrong. It takes time to elaborate on the justification of that. The auto-generated explanation behind the warning or flagging was mostly accurate, and it was like putting the logic into words. It helps to understand what you are building with the rule and what the potential results are. It was like seeing the rule in action." P10 similarly used DecideTogether's decisions from the moderation queue to identify systematic patterns of misalignment, saying that "the wrong ones were both factual statements posed as an opinion. That seemed to be the main thing that should be changed".

Refining When communicating their refined moderation intent back to the system, editing the rule text was considered the most natural way. P3 said, "writing a plain text in natural language is the easy way for me to interact with the system [other adjustments would have] a lot of learning curve." On the other hand, other features were rarely used. For instance, P8 reported that they rarely adjusted the community context after writing the initial rule, as they felt reluctant to do so based on a couple of cases. Some participants wanted more direct control over the automoderator logic, as it was the critical part controlling DecideTogether's automoderation decisions. In such cases, participants had a very specific idea of the automoderation to implement, such as a set of keywords to filter out. During the user study, the feature that suggested fixes to the logic was not commonly used by participants. Participants often agreed with most of DecideTogether's decisions or had specific ideas for editing the logic themselves, rather than using the feature.

6.6.2 RQ2: How does DecideTogether affect moderators’ understanding, trust, and control over automoderation rules?

Participants perceived the trust towards subjective automoderation with DecideTogether differently from existing automoderation tools. Participants described existing automoderation tools as keyword-based and deterministic tools. The errors with them are clear-cut and rare, requiring almost no supervision. P10 contrasted this directly with subjective tools: “Since the existing ones are only based on keywords, there is a higher sense of trust. You can have false flags, [but] it’s far less likely than subjective stuff... If it’s a keyword, it’s fine to make it take preliminary action—deterministic stuff can be, but not subjective stuff.” On the other hand, DecideTogether is capable of making moderation decisions with subjective judgments that can “keep the spirit of the community in line” (P3). However, such subjective moderations are contestable. Community members can easily get offended or disagree with the decision, which can damage the community and the authority of the moderation if the decision is not justified enough. In that sense, P3 and P10 felt that DecideTogether would require human supervision to ensure the legitimacy.

If used with human supervision, participants considered DecideTogether’s decisions trustworthy. Participants commonly mentioned two factors that contributed to their trust in DecideTogether. First, participants valued the ability to verify DecideTogether’s decisions against their own judgment, supported by the visibility of decision-level rationale. P3 said, “Seeing the logic behind the decisions is the thing that helped me to build trust quickly. The aggregated explanation was like putting in words the same logic that I would use.” The moderation queue provided the empirical ground for this verification. Interestingly, P6 perceived DecideTogether’s tendency to overflag violations as a useful design feature, ensuring that all grey-area contributions are flagged for human review. Second, DecideTogether’s support for iterating on the logic also contributed to participants’ sense of trust. While DecideTogether often gave decisions apart from what participants expected, they still believed that they could tweak the logic as much as they wanted to reach better alignment. P7 said, “(Decidetogogether) made it very clear: these are the parameters, this is what my decision was based on. It was mostly just like, oh the machine — let’s try and fix it.” These two mechanisms addressed the underlying concerns that participants like P10 expressed about AI: “You can’t be sure AI will act the way you want it to, even if the rule is clear for humans.” Beyond supporting moderators’ own trust, the visible rationale extended across stakeholder relationships. P8 saw it as a source of procedural legitimacy for community members affected by moderation: “In other communities, when you get banned, the reason is usually given as a single line. But with this system, it explains in logical detail which part violated which policy and why. There would be far fewer people who feel they were unfairly banned.” P6 extended the same transparency to coordination across moderation teams: “The decisions will be tracked... so it’s really nice to see the history. If a member ever gets upset and says ‘Hey, my post is removed,’ anyone on my team could look at the change log and see why we made that decision.”

The subjective nature of DecideTogether’s decisions also reshaped participants’ understanding of *control* over the system. In deterministic automoderation, control is primarily exercised through rule specification—configuring keywords, regular expressions, or threshold parameters that the system will then execute autonomously. In DecideTogether’s subjective context, however, participants felt the sense of control from giving the final human judgment at the point of action. P3 articulated this most directly: “As long as it doesn’t delete and ban users, I think I’m still in control.” P6, who managed a 70,000-member Facebook community, trusted DecideTogether more than Facebook’s automoderation

precisely because Facebook auto-removes posts while DecideTogether only flags them. She noted, “With Facebook... it just removes the post for you. In terms of it just being flagged, I would definitely trust it because I still make the final decision.” At the same time, participants identified a distinct dimension of control where DecideTogether fell short: *granularity*. While participants felt sufficient control over the rule as a whole, P6 and P9 wanted finer-grained means to shape individual logic items. P9 expressed wanting a per-logic-item natural-language context entry similar to the rule-level draft generation, observing that without it, “it didn’t coordinate with me just like I hoped.” P6 described intuitively wanting to edit specific logic nodes directly but found the existing interface too keyword-oriented for the kind of nuanced adjustments they had in mind. These gaps suggest that as subjective automoderation systems mature, control affordances may need to operate at multiple granularities—not only at the rule level but at the level of individual logic components, where moderators can locally shape how a single decision criterion is interpreted.

Taken together, these findings reposition control as a precondition for trust rather than a parallel concern. Participants trusted the subjective decisions of DecideTogether because subjective decisions were paired with non-autonomous execution. P1 made the causal direction explicit: “Knowing that I would be able to tweak things a lot gives me a sense of trust because it gives me a sense of control.” In deterministic automoderation, where errors are rare and clear-cut, autonomous execution may be acceptable. In subjective automoderation, where decisions are contestable and consequences are higher, the system’s willingness to defer to human authority is not a limitation to be engineered away—it is the very property that makes such systems trustworthy to deploy.

6.7 Discussion

Our findings suggest that supporting subjective moderation is not simply a matter of improving automated decisions, but of helping moderators articulate, inspect, and retain authority over the criteria that govern those decisions.

6.7.1 Intent Articulation through Reviewing Real-World Cases

A premise of DecideTogether’s design is that moderators’ enforcement intent is not fully formed in advance, but becomes legible through iteration. In that sense, we defined the end goal of DecideTogether as making a rule whose enforcement consequences the moderator has examined on concrete cases and either endorsed or corrected, rather than the well-formed rule text itself. To support the process of iterating on moderation intents and manifesting them into human-readable rules and machine-enforceable logic, DecideTogether keeps the rule text and the automoderator logic aligned. Our findings substantiate this premise and clarify the mechanism behind it. Participants did not articulate their intent by introspecting more carefully at the drafting stage; they articulated it by confronting concrete decisions and finding them wrong.

The drafting stage itself involved little friction. Most participants accepted the DecideTogether-generated rule text with minimal modification, and some valued it for surfacing considerations they had not thought of on their own (P9). Yet the substantive shaping of intent happened afterward, when participants reviewed the automoderator’s verdicts on real posts. For instance, P6 revised their rule text and the logic after seeing harmless posts flagged as misinformation, recognizing that the authors were seeking information rather than asserting a claim. P10 noticed a recurring pattern from their

initial rule, factual statements posed as opinions, only after inspecting a series of decisions. In both cases, the criterion became explicit only when a concrete case demanded it. This reframes what the moderation queue affords. Here it functioned as a medium for articulation: each verdict serves as a concrete proposition about the rule that the moderator can affirm or reject. The role of cases in drawing out the intent behind a rule echoes prior work on case-based provocation [106, 105].

At the same time, a tension stems from the fact that articulation and communication place different representational demands on moderation policies. DecideTogether treats the automoderator logic as a faithful expansion of the rule text, but the two serve different audiences. The logic addresses the machine and needs to be explicit and precise. The rule text addresses community members: it communicates what the community expects [243, 244], and moderators often leave it deliberately general, preserving moderators’ discretion and keeping enforcement criteria from being gamed [226, 245, 246]. While moderators repair the faulty logic to align with their intents by revising the rule text, it could push excessive operational detail into the rule text, whose value partly depends on omitting it. Future work can explore interaction designs that preserve alignment while accommodating the different communicative roles of rules and enforcement logic. One possibility is to introduce an intermediate representation of moderation intent, expressed in natural language, that captures the rationale that moderators develop as they review cases. Such a representation could serve as a bridge between community-facing rules and machine-executable logic, allowing each artifact to operate at an appropriate level of abstraction while remaining mutually consistent. In this way, moderators could refine enforcement criteria without necessarily embedding every operational detail into the public rule text.

6.7.2 Decomposition as a Substrate for Inspection

A recurring obstacle in LLM-based moderation is the gulf between a moderator’s intent and an opaque pipeline’s behavior: it is rarely clear which revision will produce a desired change, or whether an observed change is a genuine fix or stochastic noise [224]. DecideTogether addresses this not by making the model more accurate, but by changing the representation that moderators reason about. Rather than treating the rule as a monolithic prompt, DecideTogether decomposes it into a set of binary questions that serve as the automoderator’s internal logic.

Our findings suggest this decomposition did substantive work beyond implementation. Participants treated the logic as an account of the system’s reasoning that they could interrogate. P7 described it as the system almost explaining it back to me,” and, when a decision seemed wrong, framed the repair as locatable: there’s something faulty in the logic here—where did it go wrong?” The decomposition converts a diffuse sense that a verdict is off into a search over discrete, individually inspectable criteria. This is what narrowed the gulf of evaluation: a verdict is right or wrong because of an identifiable sub-criterion, and a revision can be aimed at that component rather than at the prompt as a whole.

It further limits the problem of LLM’s stochasticity. Stochasticity is often treated as a defect to be suppressed, and at the level of a holistic verdict, it is corrosive to trust. When decomposed into binary questions, each sub-decision is narrow enough that its outcome is comparatively stable, and the structure localizes the variation to be controlled rather than diffused across the entire decision. By making the effects of stochasticity legible and bounded, decomposition leads to more trustworthy automoderation results.

The value of decomposition may therefore extend beyond moderation. The underlying challenge is not specific to community governance but to any setting where a natural-language directive must be translated into consistent decisions. For instance, domains such as policy interpretation, content

guideline enforcement, and clinical protocol application share the structure of a natural-language directive that must be applied consistently to contested cases. Whether a decomposed, inspectable intermediate representation is a transferable strategy for such settings is, we believe, a question worth pursuing.

6.7.3 Trust and Human Authority in Subjective Automoderation

In this work, we explored automating subjective moderation. While participants generally considered DecideTogether’s moderation judgments reasonable, they commonly felt that human oversight remained necessary. Their concern was not solely whether the automoderator could make appropriate decisions, but whether subjective judgments about community norms should be delegated to an automated system in the first place.

The problem DecideTogether addresses comes into focus through the lens of street-level bureaucracy theory [230]. Alkhatib and Bernstein argue that street-level bureaucrats, such as police officers and judges, reflexively refine their decision criteria while reasoning through novel situations, whereas street-level algorithms can revise their criteria only after a decision has already been made. This “loop-and-a-half” delay yields ill-fitting decisions when novel or extenuating circumstances arise. The subjective moderation targeted by DecideTogether operates precisely in such ill-defined spaces, requiring judgments about whether content aligns with a community’s norms and values. By allowing moderators to iteratively calibrate automoderator logic against real posts and comments before deployment, DecideTogether relocates this refinement process to an anticipatory, inspectable activity while incorporating community context throughout the calibration.

Our findings, however, suggest that addressing this capability challenge does not fully resolve concerns surrounding subjective automoderation. Participants such as P3 and P10 viewed subjective moderation decisions as inherently contestable: community members may disagree with them, challenge them, or demand justification. As a result, participants questioned not only whether the automoderator could reach an appropriate judgment, but also who should hold the authority to make such judgments. Consistent with our findings on trust and control, participants trusted DecideTogether when its decisions remained reviewable and ultimately subordinate to human moderators, treating human authority as a prerequisite for trustworthy subjective automation.

This concern raises a broader question about the legitimacy of subjective automoderation. Future systems should therefore consider questions of authority and accountability alongside improvements in moderation capability. One possible direction is to make moderation criteria and automoderator logic visible and contestable not only to moderators but also to community members. For instance, P8 envisioned using DecideTogether to communicate moderation rationales and gather feedback from members about enforcement criteria. Such approaches may help subjective automoderation systems become not only more capable, but also more accountable and socially accepted.

6.8 Limitation and Future Work

This work has several limitations.

First of all, DecideTogether was evaluated in a controlled setting with a hypothetical scenario for a Reddit-like online community. While we chose to build scenarios from existing community rules to provide realistic examples, we acknowledge that problems in real-world communities may be more complex. Furthermore, different online community platforms, such as Discord, Mastodon, Fandom Wiki, and

standalone online forums, have distinct communication dynamics, platform features, and norms, leading to distinct types of community problems. We believe that future work could conduct a longitudinal evaluation of DecideTogether with real-world communities to better understand the value and limitations of DecideTogether.

Second, while we evaluated DecideTogether only in a single-user setting, community moderation is often collaborative among multiple moderators, necessitating adjustments to the moderation criteria among them [229]. We believe that DecideTogether has the potential to facilitate collaboration by externalizing the moderation intents into automoderator logics and the moderation queue. Future work could deploy DecideTogether into a real-world moderation team to better understand the potentials and the challenges of DecideTogether to facilitate communication and deliberation on community rules.

6.9 Conclusion

Subjective moderation requires a moderator to articulate a moderation intent that must serve as both a community-facing rule that members can read and a machine-facing logic that decides individual cases. Yet keeping these two forms aligned is difficult, and existing tools give moderators little visibility into how a rule becomes a decision. We built DecideTogether, a system that treats the rule text and the automoderator logic as two expressions of a single intent, compiling inspectable logic from the rule text as the source of truth and letting moderators test and refine that intent against real community cases rather than author logic directly. Our technical evaluation showed that decomposing the rule text into logic raised decision accuracy from 45.9% under direct rule-text prompting to 65.9% after error-driven updates, and that context-grounded logic was preferred in 63% of pairwise comparisons. A user study with 10 moderators showed that reviewing the system’s decisions on real cases helped them surface latent criteria and restructure vague intent into enforceable rules. With DecideTogether, we envision automated moderation that moderators can shape, inspect, and hold accountable as a faithful expression of their own moderation intent.

Chapter 7. Discussion

In this chapter, I first present reflections and considerations for designing interpretive scaffolds. Then, I discuss the potential risks and biases from interpretive scaffolds. Finally, I discuss how interpretive scaffolds could be developed further to support the social process of deliberation.

7.1 Synthesis of Empirical Support for Informed Opinion Formation

I defined an informed opinion as a position that is grounded in relevant evidence, sensitive to alternative perspectives, and supported by reasons the holder can articulate. The four studies jointly provide empirical support for this definition’s three properties: for each property, studies showed that an interpretive scaffold, placed at the moment of encounter, supported the corresponding component of informed opinion formation.

Grounding in relevant evidence requires critical consumption of information. DataDive (Chapter 4) showed that surfacing relevant datasets at the moment of reading changed what evidence entered the reader’s evaluation of a statistical claim. Participants explored significantly more external information than in the baseline condition and recalled relatively more facts beyond the article text. Sensitivity to alternative perspectives requires engagement with views unlike one’s own. StarryThoughts (Chapter 3) showed that visualizing opinions alongside their holders’ demographic identities led participants to explore opinions from a wider range of demographic groups and to read them with a more respectful attitude, thereby contributing to an expanded argument repertoire. FrameIt (Chapter 5) showed that surfacing moral foundations shifted how participants attributed disagreement, from ignorance or bad intent toward legitimate value differences. Articulate reasons require externalizing a previously implicit position. DecideTogether (Chapter 6) showed that moderators’ initially tacit intent became explicit through confrontation with concrete cases, externalized as inspectable rule text and decision logic. The argument repertoire results from StarryThoughts provides converging evidence at the individual level: after exploration, participants could state more arguments surrounding the issue, including arguments against their own position.

This evidence is component-wise. Each study examined a single component in a single encounter with a single system, but no study observed a reasoner moving across components, revising a position as engagement, consumption, and articulation feed into one another. Whether component-level support composes into a full trajectory of opinion formation is an empirical question for longitudinal work.

What the four studies do show is convergent in kind. Across different contexts, evaluating statistical claims, understanding disagreement, exploring diverse viewpoints, and articulating moderation criteria, interpretive scaffolds altered what information participants attended to, how they interpreted that information, and how they expressed the resulting judgments. This convergence suggests that supporting informed opinion formation is less a matter of supplying information in advance and more a matter of making interpretive resources available at the moment of encounter, so that information becomes evidence, perspectives become understandable, and judgments become articulable.

7.2 Design Reflections on Interpretive Scaffolds

This thesis proposes the design construct of **interpretive scaffolds**: design elements that mediate the encounter between a reasoner and a piece of information, offering alternative ways to interpret, examine, and integrate that information into the reasoner’s thinking. To support future researchers who leverage the concept of interpretive scaffolds, I distill three design decisions, each cascading from the user’s goal: the type of interpretive scaffold to use, the level of interpretive mediation to bridge the information of focus and scaffolding information, and the level of intrusiveness of user interactions.

Table 7.1: Design space for interpretive scaffolds across the four systems.

Design Axis	StarryThoughts	DataDive	FrameIt	DecideTogether
User Goal	Open-ended <i>opinion exploration</i>	<i>Critical reading</i> of statistical information	Conversation respecting <i>value</i> <i>system differences</i>	Articulating <i>community</i> <i>moderation intent</i> into rules
Scaffolding Types	Enriching	Enriching	Restructuring	Enriching + Restructuring
Scaffolding Information	Self-stated <i>demographic</i> <i>identities</i>	AI-retrieved <i>statistical data</i>	AI-analyzed <i>value</i> <i>systems</i> behind opinions	AI-generated <i>intent</i> <i>interpretation</i> + impact of intent on <i>real-world case</i>
Interaction Design	<i>Visualization of</i> <i>diversity</i>	In-text questions + <i>data visualization</i>	In-text annotation + <i>argument</i> <i>suggestion</i>	<i>Iterative workflow</i> to express and refine

7.2.1 Scaffolding Types: Should Address the Interpretive Bottleneck

The choice between enriching and restructuring should follow the designer’s diagnosis of the interpretive bottleneck users face. Existing decompositions of informed opinion formation offer a useful vocabulary here [11, 37, 247]. Some bottlenecks arise from a narrow information base: users dismiss claims without cross-validation, or remain inside algorithmic filter bubbles. The interpretive work needed is a comparison against material the user has not yet encountered, calling for enriching scaffolds. Other bottlenecks arise from shallow engagement with the information at hand: users short-circuit by dismissing an opinion or accepting a claim without examining its underlying commitments. The interpretive work needed is examination of what is already in the focal information — its assumptions, its value basis, the position it stakes out. This is what restructuring scaffolds provide.

The four systems align with different bottleneck diagnoses. DataDive addresses the evaluation of statistical claims in isolation from broader empirical context, thereby surfacing external datasets as enriching scaffolds. StarryThoughts addresses constrained exposure to opinion diversity and surfaces the demographic and ideological landscape as enriching scaffolds. FrameIt addresses the dismissal of opposing opinions on surface terms and surfaces underlying moral foundations as restructuring scaffolds.

DecideTogether addresses both at once. A moderator’s intent is undersituated in the community context, and its effect on concrete cases is not visible, so the enriching scaffold surfaces community activity and the intent’s consequences across real cases. At the same time, the criteria that would operationalize the intent remain latent, so the restructuring scaffold exposes them by translating the intent into inspectable moderation logic.

A scaffold’s mode, however, is not fixed by the resource it draws on. The same material can support either orientation, depending on how the interface positions the interpretive act. FrameIt’s moral foundations served as a restructuring scaffold when attached to a single argument, but could serve as an enriching scaffold if used to visualize the distribution of value systems across an issue. DataDive’s external datasets enrich when presented as comparative evidence, but could be restructured to expose the assumptions embedded in the focal claim. Mode, then, is a property of presentation, not of resource. And the reverse holds as well: users appropriate scaffolds beyond their intended mode. StarryThoughts’ demographic cues were designed as enriching scaffolds for opinion diversity, but participants also used them to restructure their reading of individual opinions, reflecting on assumptions about social groups.

In summary, I propose two design implications. The primary mode should align with the system designer’s diagnosis of interpretive bottleneck. But the secondary mode should remain available rather than foreclosed, because the same resource can carry both, and because users’ interpretive work routinely exceeds what designers anticipate.

7.2.2 Level of Interpretive Mediation: Match System Effort to User Capability

The level of interpretive mediation should follow the target user’s capability to connect the scaffolding information to their focal information on their own. The greater the gap between the user’s capability and the connection the scaffold requires, the heavier the mediation the system should provide.

When the user can make the connection without help, the system needs only lightweight mediation. StarryThoughts’ users read each opinion and the demographic profile attached to it directly, so the system only aggregates and visualizes the contributions and leaves the user to relate a group to its opinion distribution.

When the connection is harder, the system retrieves the scaffolding information and links it to the focal information. DataDive retrieves statistical data from outside the user’s focal context and connects it through generated questions. DecideTogether’s enriching scaffolds retrieve community activity and connect a moderator’s intent to its consequences across real cases, supplying the context the intent was undersituated in.

When the user cannot perform the interpretive operation at all, the system produces the scaffold by its own inference. FrameIt classifies the moral foundations behind an argument outright because naming and reasoning with moral foundations require a vocabulary most users lack. DecideTogether’s rule synthesis goes further in the same way, translating the latent criteria that would operationalize an intent into inspectable moderation logic.

This makes the level of mediation a derived property rather than a designer’s preference. The same source can also warrant different levels for different users: a statistical database that an expert searches through a query box may need to arrive wrapped in generated questions for a layperson.

Heavier mediation carries a cost. Each added interpretive layer raises the risk of system bias and, with LLMs, hallucination. The user is least equipped to catch these errors precisely where the system

has done the most work, since the gap that justified the heavy mediation is the same gap that limits the user’s ability to audit it. Verifiability should therefore scale with the depth of mediation, so that the more the system infers, the more it must expose the grounds of that inference for the user to check.

7.2.3 Interaction Design: from Lightweight Nudge to Deep Guidance

A third decision concerns the scaffold’s intrusiveness, or how strongly it intervenes in the user’s encounter with information. Interpretive scaffolds work by inviting deeper processing of information, which is normatively desirable but cognitively costly. Not every reasoner is willing to incur that cost on every encounter. A scaffold presented too lightly may leave the reasoners who need it most without real support, while one presented too heavily imposes loads that demotivate use altogether.

The design principle that recurs across the four systems is to open with minimal interaction and deepen it on demand, an approach that inherits the spirit of progressive disclosure in interaction design [248]. The opening move is not the scaffolding information itself but a low-cost signal that an alternative interpretation exists: an opinion distribution in *StarryThoughts*, a highlight on a statistical claim in *DataDive*, a moral foundation tag in *FrameIt*, and a community context tag in *DecideTogether*. The signal shows that the focal information is not the whole picture, and it costs almost nothing to ignore. Richer interaction follows only if the user wants it. *DataDive* guides the reader from a flagged statistic to questions that frame what comparative data would clarify; *FrameIt* from a moral foundation tag to the argument structure beneath it; *StarryThoughts* through demographic and ideological filters over the opinion landscape; and *DecideTogether* through reflection and articulation steps that move from raw intent toward a community-fit rule. The user controls how far to go.

Keeping the cost of ignoring a scaffold low is one side of motivation. The other is giving the reasoner a reason to look further, and the system suggests several strategies. One is to provoke curiosity. In *StarryThoughts*, users came in with assumptions about which demographic groups would hold which opinions, and when the actual distribution diverged from those assumptions, the mismatch drew them into closer exploration to understand why those groups thought differently. A second is to connect the information to what the reasoner personally has at stake. *DataDive* lets readers relate a statistic to their own context, such as comparing a figure to their own country, so that engaging with the claim feels directly relevant. A third is the prospect of having to articulate or justify a position to others, which encourages closer engagement in advance. *FrameIt* readers prepared responses for an ongoing conversation, and *DecideTogether* moderators wrote rules that other moderators and community members would act on. A fourth is to attach the scaffold to a task the reasoner already needs to complete, rather than supplying motivation anew, as in *DecideTogether*, where moderators arrived with the concrete task of writing rules, while open-ended reading offers no such anchor. These strategies share a common stance. They amplify the motivation the reasoner already holds rather than manufacture it, keeping the choice to engage with the user. A scaffold left at the level of a low-cost signal is easy to pass over, and I believe future work could seek more diverse strategies to motivate voluntary engagement.

7.3 Potential Biases and Risks from Interpretive Scaffolds

The concept of interpretive scaffolds was designed to prevent information consumers from being fixated on the surface meaning of the information and on their prior opinions, and to encourage the discovery of different interpretations of the given information. In that spirit, interpretive scaffolds are

designed to shed light on the broader information and opinion space while refraining from offering clear-cut conclusions. At the same time, I discuss potential risks and biases from interpretive scaffolds.

7.3.1 Framing by Design

The systems for identifying and providing interpretive scaffolds are not neutral. They need to make decisions on the axes of information space that enriching scaffolds would present, or on the latent structures that restructuring scaffolds would present. Such decisions about what to present and what not to present can frame how users interpret the information. For instance, while *StarryThoughts* used age, gender, and political identities to present the diversity of opinions, such a decision might obscure other, more important axes to explore, such as value systems or socioeconomic status. Designing interpretive scaffolds, therefore, requires making these framing decisions deliberate and, where possible, transparent, so that the presented structure reads as one selection among possible others rather than as the structure of the issue itself.

7.3.2 Bias from the AI Pipeline

Beyond the designer’s framing, the AI-powered pipelines that generate interpretive scaffolds may embed their inherent biases in their outputs, potentially amplifying the selection bias or homogenizing users’ interpretations. The risk is compounded with presentation. When a scaffold is produced by an automated pipeline and delivered through a polished interface, users may perceive it as a neutral, complete, and authoritative summary rather than as one partial framing among many. *FrameIt*’s moral foundation labels, which are over-annotated relative to human coders, illustrate the danger: a probabilistic inference about the values behind a comment can be taken as a definitive reading. This perception invites overreliance, turning the scaffold into a signal to close the interpretive space rather than open it and quietly prescribing the “right” interpretation. The mitigation is to communicate the scaffold’s partiality. Scaffolds should be presented in a provisional, revisable form that invites questioning rather than deference [249] and paired with means of audit, such as provenance for retrieved data, exposed inference steps for generated interpretations, and explanations the user can check against their own judgment.

7.3.3 Legitimation of Bad-Faith Content

A different risk arises not from the system but from how scaffolds interact with adversarial content. Several scaffolds were designed to foster charitable, respectful engagement. *FrameIt* surfaces the moral foundations beneath an argument so that opponents appear as people reasoning from different values, and *StarryThoughts* attaches demographic context to opinions so that positions are read in light of who holds them. When the underlying content is misinforming, manipulative, or extremist, the same move can lend it unearned legitimacy, presenting a claim that warrants contestation or removal as a good-faith expression of values to be understood. *FrameIt*’s deployment surfaced a milder version of this tension, in which awareness of fundamental value differences sometimes led to a premature “agreeing to disagree” rather than deeper engagement.

The corrective is not indiscriminate skepticism. Critical scrutiny is not applied evenly: people subject attitude-incongruent information to harsher examination than congenial information, a pattern of motivated skepticism [17], and skepticism applied without discrimination erodes the distinction between well-grounded and poorly-grounded claims, leaving users to fall back on prior belief as a default [250, 251]. I expect two kinds of support to be possible. One would be providing authoritative grounds against

which a factual claim can be checked, for instance, the claim’s provenance or supporting evidence from trustworthy sources such as academic literature. This gives users material to judge a claim’s standing themselves rather than directly issuing a verdict. At the same time, it presupposes users’ expertise to evaluate those grounds. The other support could be on exposing how a claim moves from its evidence to its conclusion. It surfaces the inferential steps a claim depends on, for instance, whether it generalizes from a single case, treats the absence of counterevidence as proof, or rests on a misplaced authority. A reader can spot such moves even without the expertise to evaluate the underlying facts. A growing line of systems surfaces such reasoning flaws in place as readers encounter them [252, 253]. The opportunity here is to treat this not as a standalone fallacy-detection technique but as another application of the restructuring mode that FrameIt and DecideTogether instantiate.

7.3.4 Scaffolds and Interpretive Skill

A final risk concerns the relationship between scaffolds and the user’s own interpretive skill. First, by performing interpretive work on the user’s behalf, a scaffold may prevent the development of the skill it substitutes for. For instance, a reasoner who relies on DataDive to generate comparative questions or on FrameIt to name the values behind an argument may not build the capacity to perform these operations unaided and may be left unsure how to reason about an issue when the system is absent. Second, and in the opposite direction, making sense of a scaffold can itself demand literacy, whether statistical, rhetorical, or domain-specific. Reasoners who already hold that literacy gain the most, while those without it gain the least, even though the latter are often the group a scaffold was most intended to help. A scaffold that presupposes more interpretive skill than a reasoner brings can widen rather than narrow the disparity in how well-equipped people are to reason about an issue.

Neither problem is fully resolved by guidance, since guidance presupposes a baseline on which to build. Interpretive scaffolds should therefore aim to nurture the skills they support rather than only supply their outputs, for example, by adapting the level of mediation to the user and gradually transferring interpretive work back to them. They should be evaluated not only by whether they support reasoning in aggregate, but by whether that support holds across reasoners with differing levels of relevant literacy.

7.4 Interpretive Scaffolds for the Social Process of Deliberation

This dissertation focused on supporting an individual’s analytic process to explore diverse information and arguments on social issues and to take an informed position. The contribution of this dissertation falls into what Black [11] characterizes as the *analytic* dimension of deliberation—helping people understand issues, evidence, and arguments. Yet deliberation is not solely an analytic activity. Deliberative democracy depends on a social process in which citizens exchange reasons, engage constructively across disagreement, and work toward collective judgment through reciprocal communication [9, 8, 10]. At the systemic level, deliberation further depends on the inclusion of diverse voices and perspectives throughout a broader deliberative system [254].

The systems developed in this dissertation were not designed to directly facilitate discussion between participants. Rather, they were designed to support individuals as they encounter, interpret, and reason about information. Consequently, interpretive scaffolds cannot directly resolve structural inequalities, status hierarchies, or interactional dynamics that emerge during deliberation [255, 256]. Nevertheless,

the findings of this dissertation suggest that interpretive scaffolds may support several individual-level conditions that enable productive deliberative communication. More importantly, they point toward a broader design space for technologies that support deliberation not by managing discussion itself, but by shaping how participants make sense of information, disagreement, and one another before and during deliberative encounters.

Below, I discuss how interpretive scaffolds may help support the components of the social process of deliberation: respectful engagement across disagreement, mutual understanding, inclusive and balanced participation, and the development of a shared basis for collective judgment.

7.4.1 Respectful Engagement Across Disagreement

A foundational norm of deliberation is that participants treat one another as legitimate contributors despite disagreement [9]. This norm is difficult to sustain because disagreement is often framed in terms of social categories rather than reasons. Under affective polarization, opponents are perceived not as people reasoning from different premises but as members of distrusted groups [257, 258, 259].

The opportunity for interpretive scaffolding here is to shift what a participant attends to when disagreement is encountered. Rather than supplying more arguments, a scaffold can surface the values, experiences, and assumptions that make an opposing position intelligible, so that disagreement is read as reasoning from different premises rather than as ignorance or bad faith [260, 261, 118]. This is an interpretive intervention on the perception of the opponent, not on the issue’s content. Findings from FrameIt and from the cross-cutting identities in StarryThoughts offer preliminary evidence that such reframing is achievable.

7.4.2 Mutual Understanding

Deliberation requires participants to understand not only what others believe but why they hold those beliefs [9]. Failures of understanding often arise not because information is unavailable but because the interpretive context needed to make sense of a position remains implicit. The assumptions, value conflicts, and conditions under which a view holds are rarely visible alongside the view itself.

Interpretive scaffolding suggests a way to address this without resolving the disagreement. A scaffold can make interpretive context available at the moment of reasoning by surfacing the assumptions and reasoning structures that support a competing position, so that a participant grasps the pathway to a conclusion and not only the conclusion. FrameIt, by surfacing the moral foundations underlying a position, and DecideTogether, by making the reasoning behind a judgment explicit, offer preliminary evidence that this pathway can be made legible to someone who does not share the conclusion.

7.4.3 Inclusive Participation

Deliberative systems depend on the inclusion of diverse perspectives [254], yet many perspectives never enter discussion. Individuals who wrongly believe their view is held by a small minority stay silent [262, 133], and others withdraw because they anticipate conflict or social sanction [263].

Both failures are interpretive before they are behavioral: they rest on a participant’s reading of the distribution of opinion and of the social risk of speaking. A scaffold can act on that reading by making otherwise hidden perspectives visible and by lowering the perceived risk of expressing disagreement, so that a participant comes to see their own view as a legitimate contribution rather than a deviant one.

StarryThoughts, by making the distribution of viewpoints observable, and FrameIt, by reducing the expectation of attack, provide preliminary evidence for this.

7.4.4 A Shared Basis for Collective Judgment

Productive deliberation requires shared informational grounding. Participants need not agree on conclusions, but they must be able to understand the evidence and reasoning behind competing judgments [264, 10]. Much public discourse is organized around conclusions rather than reasoning, so disagreement settles on outcomes while the underlying assumptions stay opaque.

Interpretive scaffolding addresses this by making reasoning itself a shared object. A scaffold can externalize the assumptions, trade-offs, and decision logic behind a judgment, allowing participants to deliberate over the interpretive structures that produced the outcome rather than just the outcome itself. DecideTogether (Chapter 6) illustrates this by linking judgments to the rules and reasoning that generated them, making them inspectable and contestable. Such shared reasoning is one route to constructing shared realities that support collective judgment.

7.4.5 Summary

Much work on deliberation technology connects people to information, arguments, or discussion partners. Interpretive scaffolding points toward a complementary goal: helping participants construct the interpretive context needed to make sense of what they encounter and whom they encounter. Across the four preconditions above, the intervention is the same in kind. It does not direct participants toward conclusions but makes latent perspectives, assumptions, and reasoning available when needed. This dissertation offers an initial exploration of that possibility and suggests interpretive scaffolding as a design paradigm to support the social process of deliberation, rather than merely as an analytic foundation.

7.5 Toward User-Driven Interpretive Scaffolds

A common thread runs through these systems and through the risks discussed above. In every case, it is the system that decides what to illuminate: DataDive decides what counts as a relevant comparison, FrameIt what counts as a meaningful frame, StarryThoughts which identity axes count as salient, and DecideTogether which cases count as instructive. This is what makes the scaffolds tractable, and it is also what concentrates interpretive authority in the system. The framing introduced by design, the bias introduced by the pipeline, and the question of who provides the scaffold are all consequences of this concentration.

A natural direction for future work is to move interpretive authority toward the user. Rather than receiving a fixed set of illuminated dimensions, users would aim their own spotlights, choosing which comparisons, frames, or axes to surface, and would contribute to the scaffolding resource itself rather than only consuming it. StarryThoughts already gestures at this, since its opinion landscape is assembled from user contributions rather than retrieved from an external source. Under a user-driven model, the system’s role shifts from deciding what is worth seeing to helping users decide for themselves and pool what they surface, thereby redistributing the framing decisions that would otherwise rest with the designer. Together with the extension from the analytic to the social dimension of deliberation, such an extension would enable interpretive scaffolds to foster interpersonal communication and learning.

7.6 Limitations

Nature of the evidence. The claims in this dissertation rest largely on self-report and proxy measures collected over short study periods. StarryThoughts measured change through argument repertoire and self-reported opinion strength, and the other systems relied on perceived change, interview accounts, and short-term task performance. These measures indicate that the scaffolds expanded the interpretive resources participants engaged with, but they do not establish that participants formed more informed or better-reasoned opinions in a behavioral sense, nor that any effect persisted beyond the study. Whether interpretive scaffolds change how people actually reason about contested issues over time remains an open empirical question.

Scope of the evaluations and the skill claim. Each system was evaluated on a small set of contested issues with specific participant populations and under short exposure. This limits how far the construct generalizes across issues, cultures, and platforms. The limitation is sharper for the claim that scaffolds can nurture interpretive skill rather than substitute for it. No study tracked whether participants retained or transferred interpretive practices after the scaffold was removed, so the dissertation does not establish that scaffolds build durable capability rather than producing effects that depend on the scaffold’s presence.

Residual bias from AI mediation. The verifiability strategies introduced in this work, such as exposing the underlying data in DataDive and surfacing the reasoning behind the moral foundation inference in FrameIt, reduce, but do not remove, the bias introduced by the AI pipeline. Deeper interpretive mediation improves contextual fit but also increases the influence of what the pipeline chooses to retrieve or generate. The dissertation did not systematically audit the distribution of interpretations these pipelines surface, so the extent to which they over- or under-represent parts of the interpretive space is not characterized.

Chapter 8. Conclusion

This dissertation introduced the concept of interpretive scaffold, which is a design element mediating between a reasoner and a piece of information, showing alternative ways to interpret, examine, and integrate that information into reasoning, enabling users to form informed opinions on social issues. To conclude, this chapter summarizes the main contributions of the thesis and proposes directions for future research.

8.1 Summary of Contributions

This dissertation makes three contributions to the study of how interactive systems can support informed opinion formation on contested social issues.

- **Conceptual: the interpretive scaffold construct.** I introduce interpretive scaffolds as a design construct for supporting individual reasoning at the moment of encountering a piece of information by showing alternative ways to interpret, examine, and integrate the information. I motivate the construct from three challenges to informed opinion formation, and I distinguish two modes, enriching and restructuring, that target different challenges. The construct serves as a lens for organizing the design of the systems in this dissertation, and I build on it to propose a set of design considerations that traces three design decisions from a user’s goal, offered to guide future work on systems for reasoning support.
- **Artifact: systems and techniques.** I contribute design strategies and computational techniques for realizing interpretive scaffolds in practice. Across the four systems, StarryThoughts, DataDive, FrameIt, and DecideTogether, I demonstrate how interpretive support can be elicited from external sources and from the information of focus itself, surfaced at the moment of reading without disrupting the reasoner’s process, and integrated with the reasoner’s ongoing thoughts. The four systems show that interpretive scaffolds can draw on resources of different kinds and operate at different depths of interpretive mediation.
- **Empirical: evidence from user studies.** I provide evidence from user studies of each system, characterizing how interpretive scaffolds shape reasoning across four tasks: engaging with diverse opinions, assessing statistical claims, examining the value basis of arguments, and translating abstract moderation intent into rules. I synthesize these findings into an account of how interactive systems can support individual analytic work on contested social issues.

8.2 Future Directions

I discuss several directions for future research.

8.2.1 Longitudinal Evaluation of Interpretive Scaffolds

The limitations in Section 7.6 point to a common gap: the evaluations in this dissertation captured effects during or immediately after short exposure, leaving open whether those effects persist and

whether scaffolds build durable interpretive capability. A longitudinal study would address this directly. Deploying the systems over an extended period, for instance as a browser add-on encountered during everyday reading, would allow observation of how repeated use shapes the way people approach contested information over time. Such a study could examine whether interpretive practices transfer to settings where the scaffold is absent, distinguishing genuine skill development from effects that depend on the scaffold being present, and whether any change in how people form and revise opinions holds beyond the study period.

8.2.2 Interpretive Scaffolds for Agent-Mediated Information Consumption

People increasingly obtain information by asking conversational LLM agents rather than searching or reading primary sources. The agent retrieves and synthesizes across sources and returns a composed answer, and that answer becomes the surface on which the user forms an opinion. Interpretive scaffolds could be applied at this surface, treating the agent’s response as the information being encountered and layering interpretive resources onto it. This raises a design problem distinct from the prior systems: the surface is generated on demand and varies from query to query, rather than a persistent text or visualization that can be annotated in advance.

8.2.3 Multimodal Information

Information about social issues now arrives substantially through multiple modalities, such as videos, live-streamings, short-form clips, and podcasts. The designs of interpretive scaffolds I presented assume a persistent textual or visual surface to annotate so the user can explore diverse interpretations, but such designs may not transfer to audio-only or time-based channels where attention is serialized by playback. Surfacing interpretive resources in these channels, for example by layering commentary into playback or adding interactive elements on videos, is an open design problem.

Bibliography

- [1] Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L. Freed. Effective messages in vaccine promotion: a randomized trial. *Pediatrics*, 133(4):e835–842, April 2014.
- [2] Dan M. Kahan, Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel. The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature climate change*, 2(10):732–735, 2012.
- [3] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale.ORIM, June 2018.
- [4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, July 2016. arXiv:1606.06565 [cs.AI].
- [5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, pages 610–623, New York, NY, USA, 2021. Association for Computing Machinery.
- [6] Jens Hainmueller and Daniel J. Hopkins. Public Attitudes Toward Immigration. *Annual Review of Political Science*, 17(Volume 17, 2014):225–249, May 2014.
- [7] Michael J. Sandel. *Justice: What’s the Right Thing to Do?* Farrar, Straus and Giroux, New York, NY, 2010.
- [8] Juergen Habermas. *The Theory of Communicative Action: Volume 1: Reason and the Rationalization of Society*. Beacon Press, 1984.
- [9] Amy Gutmann and Dennis F Thompson. *Why deliberative democracy?* Princeton University Press, 2004.
- [10] Joshua Cohen. Deliberation and democratic legitimacy. In Alan Hamlin and Philip Pettit, editors, *The Good Polity: Normative Analysis of the State*, pages 17–34. Basil Blackwell, Oxford, 1989.
- [11] Laura W. Black, Howard T. Welser, Dan Cosley, and Jocelyn M. DeGroot. Self-Governance Through Group Discussion in Wikipedia: Measuring Deliberation in Online Groups. *Small Group Research*, 42(5):595–634, October 2011.
- [12] Robert E. Goodin. Democratic Deliberation Within. *Philosophy & Public Affairs*, 29(1):81–109, 2000.
- [13] Robert E. Goodin and Simon J. Niemeyer. When Does Deliberation Begin? Internal Reflection versus Public Discussion in Deliberative Democracy. *Political Studies*, 51(4):627–649, December 2003.
- [14] Jonathan St B. T. Evans and Keith E. Stanovich. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 8(3):223–241, May 2013.

- [15] Arthur Lupia. *Uninformed: Why People Seem to Know So Little about Politics and What We Can Do about It*. Oxford University Press, New York, 2016.
- [16] Ziva Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990.
- [17] Charles S. Taber and Milton Lodge. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, 50(3):755–769, 2006.
- [18] Dan M. Kahan. Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, 8(4):407–424, July 2013.
- [19] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.
- [20] Hugo Mercier and Dan Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74, April 2011.
- [21] Deanna Kuhn. *The skills of argument*. Cambridge University Press, 1991.
- [22] D. N. Perkins. Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, 77(5):562–571, 1985.
- [23] Robert M. Entman. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58, 1993.
- [24] Dennis Chong and James N. Druckman. Framing Theory. *Annual Review of Political Science*, 10(Volume 10, 2007):103–126, June 2007.
- [25] Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, September 1974.
- [26] Daniel Kahneman and Dale T. Miller. Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93(2):136–153, 1986.
- [27] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- [28] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, March 2018.
- [29] M. J. Crockett. Moral outrage in the digital age. *Nature Human Behaviour*, 1(11):769–771, November 2017.
- [30] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.
- [31] John S. Dryzek. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford University Press, January 2002.
- [32] Hélène Landemore. Democratic Reason: The Mechanisms of Collective Intelligence in Politics. In Hélène Landemore and Jon Elster, editors, *Collective Wisdom: Principles and Mechanisms*, pages 251–289. Cambridge University Press, Cambridge, 2012.

- [33] Russell Hardin. Deliberative democracy. *Contemporary debates in political philosophy*, pages 229–246, 2009.
- [34] Engin Bozdag and Jeroen van den Hoven. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265, December 2015.
- [35] Robert E. Goodin. *Reflective democracy*. Oxford University Press, Oxford, 2003.
- [36] Raymond J. Pingree. How Messages Affect Their Senders: A More General Model of Message Effects and Implications for Deliberation. *Communication Theory*, 17(4):439–461, November 2007.
- [37] James S Fishkin and Robert C Luskin. Experimenting with a Democratic Ideal: Deliberative Polling and Public Opinion. *Acta Politica*, 40(3):284–298, 2005.
- [38] Arthur Lupia and Mathew D. McCubbins. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge University Press, March 1998.
- [39] John R. Zaller. *The Nature and Origins of Mass Opinion*. Cambridge University Press, 1 edition, August 1992.
- [40] Horst W. J. Rittel and Melvin M. Webber. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169, June 1973.
- [41] David H. Jonassen. Instructional design models for well-structured and III-structured problem-solving learning outcomes. *Educational Technology Research and Development*, 45(1):65–94, March 1997.
- [42] Jennifer Wiley and James F. Voss. Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, 91(2):301–311, 1999.
- [43] Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. Routledge, London, March 2017.
- [44] Philip M. Fernbach, Todd Rogers, Craig R. Fox, and Steven A. Sloman. Political extremism is supported by an illusion of understanding. *Psychological Science*, 24(6):939–946, June 2013.
- [45] Leonid Rozenblit and Frank Keil. The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26(5):521–562, September 2002.
- [46] E. Pariser. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited, 2011.
- [47] Cass R. Sunstein. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press, Princeton, 2017.
- [48] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, June 2015.
- [49] Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320, January 2016.

- [50] Andrew Guess, Brendan Nyhan, and Jason Reifler. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*, 9(3):4, 2018.
- [51] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1175–1184, New York, NY, USA, 2010. Association for Computing Machinery.
- [52] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 443–452, New York, NY, USA, 2009. Association for Computing Machinery.
- [53] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 265–274, New York, NY, USA, 2012. Association for Computing Machinery.
- [54] Sean Munson, Stephanie Lee, and Paul Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM '13, pages 419–428, 2013.
- [55] Mingkun Gao, Hyo Jin Do, and Wai-Tat Fu. Burst Your Bubble! An Intelligent System for Improving Awareness of Diverse Social Opinions. In *23rd International Conference on Intelligent User Interfaces*, pages 371–383, Tokyo Japan, March 2018. ACM.
- [56] Bryan Semaan, Heather Faucett, Scott P. Robertson, Misa Maruyama, and Sara Douglas. Designing Political Deliberation Environments to Support Interactions in the Public Sphere. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, pages 3167–3176, 2015.
- [57] Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey Hancock, Jeanne L. Tsai, and Michael S. Bernstein. Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, 390(6776):eadu5584, 2025.
- [58] Chenyan Jia, Michelle S. Lam, Minh Chau Mai, Jeffrey T. Hancock, and Michael S. Bernstein. Embedding Democratic Values into Social Media AIs via Societal Objective Functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):163:1–163:36, 2024.
- [59] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 81–90, New York, NY, USA, 2017. Association for Computing Machinery.
- [60] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation, October 2022. arXiv:2210.15723 [cs].

- [61] Farnaz Jahanbakhsh, Dora Zhao, Tiziano Piccardi, Zachary Robertson, Ziv Epstein, Sanmi Koyejo, and Michael S. Bernstein. Value alignment of social media ranking algorithms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery.
- [62] Juho Kim, Eun-Young Ko, Jonghyuk Jung, Chang Won Lee, Nam Wook Kim, and Jihee Kim. Factful: Engaging taxpayers in the public discussion of a government budget. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2843–2852, New York, NY, USA, 2015. Association for Computing Machinery.
- [63] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. Contextifier: Automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2707–2716, New York, NY, USA, 2013. Association for Computing Machinery.
- [64] Tong Gao, Jessica R. Hullman, Eytan Adar, Brent Hecht, and Nicholas Diakopoulos. Newsviews: An automated pipeline for creating custom geovisualizations for news. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 3005–3014, New York, NY, USA, 2014. Association for Computing Machinery.
- [65] Allen Yilun Lin, Joshua Ford, Eytan Adar, and Brent Hecht. Vizbywiki: Mining data visualizations from the web to enrich news articles. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 873–882, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.
- [66] Levy Silva and Luciano Barbosa. Matching news articles and wikipedia tables for news augmentation. *Knowledge and Information Systems*, 65(4):1713–1734, April 2023.
- [67] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.
- [68] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Trans. Comput.-Hum. Interact.*, 30(5), September 2023.
- [69] Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. In *27th International Conference on Intelligent User Interfaces*, IUI '22, pages 707–719, New York, NY, USA, 2022. Association for Computing Machinery.
- [70] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, pages 1–15, New York, NY, USA, 2022. Association for Computing Machinery.
- [71] Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. Generating personalized spatial analogies for distances and areas. In *Proceedings of the 2016 CHI Conference on Human Factors in*

- Computing Systems*, CHI '16, page 38–48, New York, NY, USA, 2016. Association for Computing Machinery.
- [72] Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. Improving comprehension of measurements using concrete re-expression strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.
 - [73] Yanan Wang and Yea-Seul Kim. Making data-driven articles more accessible: An active preference learning approach to data fact personalization. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, DIS '23, page 1353–1366, New York, NY, USA, 2023. Association for Computing Machinery.
 - [74] Pablo J. Barrio, Daniel G. Goldstein, and Jake M. Hofman. Improving Comprehension of Numbers in the News. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2729–2739, New York, NY, USA, 2016. Association for Computing Machinery.
 - [75] Martin Saveski, Nabeel Gillani, Ann Yuan, Prashanth Vijayaraghavan, and Deb Roy. Perspective-Taking to Reduce Affective Polarization on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:885–895, May 2022.
 - [76] Matti Nelimarkka, Jean Philippe Rancy, Jennifer Grygiel, and Bryan Semaan. (Re)Design to Mitigate Political Polarization: Reflecting Habermas’ ideal communication space in the United States of America and Finland. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, November 2019.
 - [77] Andrea Mauri, Yen-Chia Hsu, Himanshu Verma, Andrea Tocchetti, Marco Brambilla, and Alessandro Bozzon. Policy Sandboxing: Empathy As An Enabler Towards Inclusive Policy-Making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):369:1–369:42, 2024.
 - [78] Hyunwoo Kim, Eun-Young Ko, Donghoon Han, Sung-Chul Lee, Simon T. Perrault, Jihee Kim, and Juho Kim. Crowdsourcing Perspectives on Public Policy from Stakeholders. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, Glasgow Scotland Uk, May 2019. ACM.
 - [79] Q. Vera Liao and Wai-Tat Fu. Can you hear me now? mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 184–196, New York, NY, USA, 2014. Association for Computing Machinery.
 - [80] Q. Vera Liao and Wai-Tat Fu. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2745–2754, New York, NY, USA, 2014. Association for Computing Machinery.
 - [81] Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. Thread with caution: Proactively helping users assess and deescalate tension in their online discussions. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), November 2022.

- [82] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. Charagraph: Interactive Generation of Charts for Realtime Annotation of Data-Rich Paragraphs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–18, New York, NY, USA, 2023. Association for Computing Machinery.
- [83] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. Elastic documents: Coupling text and tables through contextual visualizations for enhanced document reading. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):661–671, 2019.
- [84] Tae Soo Kim, Matt Latzke, Jonathan Bragg, Amy X. Zhang, and Joseph Chee Chang. Papeos: Augmenting research papers with talk videos. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [85] Shahid Latif, Zheng Zhou, Yoon Kim, Fabian Beck, and Nam Wook Kim. Kori: Interactive synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):184–194, 2022.
- [86] Zhutian Chen and Haijun Xia. CrossData: Leveraging Text-Data Connections for Authoring Data Documents. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–15, New York, NY, USA, 2022. Association for Computing Machinery.
- [87] Matthew Conlen, Megan Vo, Alan Tan, and Jeffrey Heer. Idyll studio: A structured editor for authoring interactive & data-driven articles. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, page 1–12, New York, NY, USA, 2021. Association for Computing Machinery.
- [88] Thimo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. Al: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.
- [89] Anna De Liddo and Simon Buckingham Shum. Cohere: A prototype for contested collective intelligence. *ACM Computer Supported Cooperative Work (CSCW 2010): Workshop: Collective Intelligence in Organizations – Toward a Research Agenda*, 2010.
- [90] Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. Persua: A Visual Interactive System to Enhance the Persuasiveness of Arguments in Online Discussion. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):319:1–319:30, 2022.
- [91] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pages 1–30, New York, NY, USA, 2023. Association for Computing Machinery.
- [92] Eric P. S. Baumer, Jaime Snyder, and Geri K. Gay. Interpretive impacts of text visualization: Mitigating political framing effects. *ACM Transactions on Computer-Human Interaction*, 25(4), August 2018.

- [93] Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. Critrainer: An adaptive training tool for critical paper reading. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [94] Zhenhui Peng, Yuzhi Liu, Hanqi Zhou, Zuyu Xu, and Xiaojuan Ma. CReBot: Exploring interactive question prompts for critical paper reading. *International Journal of Human-Computer Studies*, 167:102898, November 2022.
- [95] Eunyoung Ko, Yeonsu Kim, and Juho Kim. Reviewaid: A scaffolded approach to supporting readers' evaluation of health news. In *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022*, pp. 313-320. International Society of the Learning Sciences, 2022.
- [96] Xiang "Anthony" Chen, Chien-Sheng Wu, Lidiya Murakhovs'ka, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. Marvista: Exploring the design of a human-ai collaborative news reading tool. *ACM Transactions on Computer-Human Interaction*, 30(6), September 2023.
- [97] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–16, New York, NY, USA, 2023. Association for Computing Machinery.
- [98] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–24, New York, NY, USA, 2024. Association for Computing Machinery.
- [99] Gionnieve Lim, Juho Kim, and Simon Perrault. Iffy-or-not: Critically evaluating potential misinformation using fallacy detection and socratic questioning with llms. *ACM Transactions on Computer-Human Interaction*, 33(1), February 2026.
- [100] Jessica Z. Wang, Amy X. Zhang, and David R. Karger. Designing for Engaging with News using Moral Framing towards Bridging Ideological Divides. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):42:1–42:23, 2022.
- [101] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [102] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pages 1–20, New York, NY, USA, 2023. Association for Computing Machinery.
- [103] Jeongeon Park, Bryan Min, Kihoon Son, Jean Y Song, Xiaojuan Ma, and Juho Kim. Choice-mates: Supporting unfamiliar online decision-making with multi-agent conversational interactions. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, IUI '26, page 1526–1550, New York, NY, USA, 2026. Association for Computing Machinery.

- [104] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–26, New York, NY, USA, 2024. Association for Computing Machinery.
- [105] Tzu-Sheng Kuo, Quan Ze Chen, Amy X. Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. Polycraft: Supporting collaborative and participatory policy design through case-grounded deliberation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.
- [106] Tzu-Sheng Kuo, Sophia Liu, Quan Ze Chen, Joseph Seering, Amy X. Zhang, Haiyi Zhu, and Kenneth Holstein. Botender: Supporting Communities in Collaboratively Designing AI Agents through Case-Based Provocations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, pages 1–31, New York, NY, USA, 2026. Association for Computing Machinery.
- [107] Savvas Petridis, Benjamin D Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J Cai, and Michael Terry. ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback into Principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 853–868, New York, NY, USA, 2024. Association for Computing Machinery.
- [108] Michelle S. Lam, Fred Hohman, Dominik Moritz, Jeffrey P Bigham, Kenneth Holstein, and Mary Beth Kery. Policy Maps: Tools for Guiding the Unbounded Space of LLM Behaviors. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, UIST '25, pages 1–24, New York, NY, USA, 2025. Association for Computing Machinery.
- [109] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.
- [110] Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA, 2024. Association for Computing Machinery.
- [111] Hyunwoo Kim, Haesoo Kim, Kyung Je Jo, and Juho Kim. StarryThoughts: Facilitating Diverse Opinion Exploration on Social Issues. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–29, April 2021.
- [112] Lincoln Dahlberg. The Internet and Democratic Discourse: Exploring The Prospects of Online Deliberative Forums Extending the Public Sphere. *Information, Communication & Society*, 4(4):615–633, January 2001.
- [113] Tanja Aitamurto and Hélène Landemore. Five design principles for crowdsourced policymaking: assessing the case of crowdsourced off road traffic law in Finland. *Journal of Social Media for Organizations*, 2(1):1–19, 2015.

- [114] Amy X. Zhang and Scott Counts. Modeling ideology and predicting policy change with social media: Case of same-sex marriage. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2603–2612, New York, NY, USA, 2015. Association for Computing Machinery.
- [115] Cynthia R Farina, Mary Newhart, and Josiah Heidt. Rulemaking vs. democracy: Judging and nudging public participation that counts. *Michigan Journal of Environmental and Administrative Law*, 2(1):12–20, 2012.
- [116] D.R. Forsyth. *Group Dynamics*. Cengage Learning, 2018.
- [117] C.R. Sunstein. *Republic.com*. Republic.com. Princeton University Press, 2001.
- [118] Lincoln Dahlberg. Computer-Mediated Communication and the Public Sphere: a Critical Analysis. *Journal of Computer-Mediated Communication*, 7(1), 10 2001. JCMC714.
- [119] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Amy J. Ko. Is this what you meant? promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1559–1568, New York, NY, USA, 2012. Association for Computing Machinery.
- [120] Markus Brauer and Abdelatif Er-rafiy. Increasing perceived variability reduces prejudice and discrimination. *Journal of Experimental Social Psychology*, 47(5):871–881, 2011.
- [121] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. Explaining the gap: Visualizing one’s predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1375–1386, New York, NY, USA, 2017. Association for Computing Machinery.
- [122] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017.
- [123] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [124] Kyung-min Lee. Korea divided over deferring retirement age. *The Korea Times*, Feb 2020. Available at: https://www.koreatimes.co.kr/www/biz/2020/02/367_283528.html (Accessed: June 17th, 2026).
- [125] Sam Kim. South korea’s bid to raise the retirement age faces rocky path. *Bloomberg*, Oct 2019. Available at: <https://www.bloomberg.com/news/articles/2019-10-14/south-korea-s-bid-to-raise-the-retirement-age-faces-rocky-path>, (Accessed: June 17th, 2026).
- [126] Victoria Kim. He played for 72 hours straight: South korea wrestles with video game addiction. *Los Angeles Times*, Oct 2019. Available at: <https://www.latimes.com/world-nation/story/2019-10-17/south-korea-video-game-addiction-mental-health> (Accessed: June 17th, 2026).

- [127] Eun-ji Bahk. S. Korea to discuss 'gaming disorder' with game industry, May 2019. Available at: https://www.koreatimes.co.kr/www/nation/2019/05/119_269501.html (Accessed: June 17th, 2026).
- [128] Joseph N. Cappella, Vincent Price, and Lilach Nir. Argument Repertoire as a Reliable and Valid Measure of Opinion Quality: Electronic Dialogue During Campaign 2000. *Political Communication*, 19(1):73–93, 2002.
- [129] Richard van der Wurff, Knut De Swert, and Sophie Lecheler. News Quality and Public Opinion: The Impact of Deliberative Quality of News Media on Citizens' Argument Repertoire. *International Journal of Public Opinion Research*, 30(2):233–256, 08 2016.
- [130] Tenelle Porter and Karina Schumann. Intellectual humility and openness to the opposing view. *Self and Identity*, 17(2):139–162, 2018.
- [131] Zoeë Richards and Miles Hewstone. Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, 5(1):52–73, 2001.
- [132] Bernadette Park, Carey S. Ryan, and Charles M. Judd. Role of Meaningful Subgroups in Explaining Differences in Perceived Variability for In-Groups and Out-Groups. *Journal of Personality and Social Psychology*, 63(4):553–567, 1992.
- [133] Elisabeth Noelle-Neumann. The spiral of silence a theory of public opinion. *Journal of Communication*, 24(2):43–51, 1974.
- [134] Andy Cockburn, Amy Karlson, and Benjamin B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys*, 41(1), January 2009.
- [135] Hyunwoo Kim, Khanh Duy Le, Gionnieve Lim, Dae Hyun Kim, Yoo Jin Hong, and Juho Kim. DataDive: Supporting Readers' Contextualization of Statistical Statements with Data Exploration. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 623–639, New York, NY, USA, 2024. Association for Computing Machinery.
- [136] Jessica Hullman, Nicholas Diakopoulos, Elaheh Momeni, and Eytan Adar. Content, Context, and Critique: Commenting on a Data Visualization Blog. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1170–1175, Vancouver BC Canada, February 2015. ACM.
- [137] Ben Shneiderman. Data Visualization's Breakthrough Moment in the COVID-19 Crisis, April 2020. Available from: <https://medium.com/nightingale/data-visualizations-breakthrough-moment-in-the-covid-19-crisis-ce46627c7db5> (Accessed: June 18th, 2026).
- [138] Josh Radinsky and Iris Tabak. Data practices during COVID: Everyday sensemaking in a high-stakes information ecology. *British Journal of Educational Technology*, 53(5):1221–1243, 2022.
- [139] Sara Doan. Misrepresenting COVID-19: Lying With Charts During the Second Golden Age of Data Design. *Journal of Business and Technical Communication*, 35(1):73–79, January 2021. Publisher: SAGE Publications Inc.

- [140] Tobias Kauer, Marian Dörk, Arran L. Ridley, and Benjamin Bach. The Public Life of Data: Investigating Reactions to Visualizations on Reddit. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–12, New York, NY, USA, May 2021. Association for Computing Machinery.
- [141] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. Storytelling in information visualizations: Does it engage users to explore data? In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1449–1458, New York, NY, USA, 2015. Association for Computing Machinery.
- [142] Ashley Ahn. South Korea has the world’s lowest fertility rate, a struggle with lessons for us all. *NPR*, March 2023.
- [143] Jeffrey Heer, Fernanda B. Viégas, and Martin Wattenberg. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, pages 1029–1038, San Jose, California, USA, 2007. ACM Press.
- [144] Brian James McInnis, Lu Sun, Jungwon Shin, and Steven P. Dow. Rare, but Valuable: Understanding Data-centered Talk in News Website Comment Sections. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27, October 2020.
- [145] Pertti Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, 2016.
- [146] Shagun Jhaver, Pranil Vora, and Amy Bruckman. Designing for Civil Conversations: Lessons Learned from ChangeMyView. *Georgia Institute of Technology*, page 11, 2017.
- [147] Leilani Battle, Peitong Duan, Zachery Miranda, Dana Mukusheva, Remco Chang, and Michael Stonebraker. Beagle: Automated Extraction and Interpretation of Visualizations from the Web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–8, Montreal QC Canada, April 2018. ACM.
- [148] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, January 2017. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- [149] Wibke Weber, Martin Engebretsen, and Helen Kennedy. Data stories. Rethinking journalistic storytelling in the context of data journalism. *Studies in Communication Sciences*, 18(1):191–206–191–206, November 2018. Number: 1.
- [150] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. *The Data Journalism Handbook*. O’Reilly Media, Inc., 2012.
- [151] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–15, New York, NY, USA, May 2019. Association for Computing Machinery.

- [152] Oana Balalau, Simon Ebel, Théo Galizzi, Ioana Manolescu, Quentin Massonnat, Antoine Deiana, Emilie Gautreau, Antoine Krempf, Thomas Pontillon, Gérald Roux, and Joanna Yakin. Statistical claim checking: Statcheck in action. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 4798–4802, New York, NY, USA, 2022. Association for Computing Machinery.
- [153] Hannah Ritchie, Max Roser, and Pablo Rosado. *CO₂ and greenhouse gas emissions*. *Our World in Data*, 2020. Available at: <https://ourworldindata.org/co2-and-greenhouse-gas-emissions> (Accessed: June 17th, 2026).
- [154] Hannah Ritchie, Max Roser, and Pablo Rosado. Energy. *Our World in Data*, 2022. Available at: <https://ourworldindata.org/energy> (Accessed: June 17th, 2026).
- [155] The World Bank. World development indicators, 2023. Available at: <https://datatopics.worldbank.org/world-development-indicators/> (Accessed: June 17th, 2026).
- [156] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pages 365–377, New York, NY, USA, October 2016. Association for Computing Machinery.
- [157] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, UIST '15, pages 489–500, New York, NY, USA, November 2015. Association for Computing Machinery.
- [158] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [159] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [160] Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Binding language models in symbolic languages. *ICLR*, 2023.
- [161] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, aug 2017.
- [162] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang

- Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, October 2023. arXiv:2211.09110 [cs.CL].
- [163] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.
- [164] Patrick W Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. *Usability evaluation in industry*. CRC Press, 1996.
- [165] Tarfah Alrashed, Lea Verou, and David Karger. Wikxhibit: Using HTML and Wikidata to Author Applications that Link Data Across the Web. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, pages 1–15, New York, NY, USA, 2022. Association for Computing Machinery.
- [166] Nicholas Kong and Maneesh Agrawala. Graphical overlays: Using layered elements to aid chart reading. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2631–2638, 2012.
- [167] International Monetary Fund. The New Economics of Fertility, 2022. Available at: <https://www.imf.org/en/Publications/fandd/issues/Series/Analytical-Series/new-economics-of-fertility-doecke-hannusch-kindermann-tertilt> (Accessed: June 17th, 2026).
- [168] Ian G. Johnson, Aare Puussaari, Jennifer Manuel, and Peter Wright. Neighbourhood Data: Exploring the Role of Open Data in Locally Devolved Policymaking Processes. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–20, November 2018.
- [169] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, October 2011. Publisher: SAGE Publications.
- [170] Amy X. Zhang, Michael Muller, and Dakuo Wang. How do Data Science Workers Collaborate? Roles, Workflows, and Tools, April 2020. arXiv:2001.06684 [cs.HC].
- [171] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.
- [172] Kristi Morton, Magdalena Balazinska, Dan Grossman, and Jock Mackinlay. Support the data enthusiast: challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment*, 7(6):453–456, February 2014.

- [173] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120, October 2023.
- [174] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. Social Media Is Polarized, Social Media Is Polarized: Towards a New Design Agenda for Mitigating Polarization. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 957–970, Hong Kong China, June 2018. ACM.
- [175] Michael Yeomans, Julia Minson, Hanne Collins, Frances Chen, and Francesca Gino. Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, 160:131–148, September 2020.
- [176] Julia A. Minson and Frances S. Chen. Receptiveness to Opposing Views: Conceptualization and Integrative Review. *Personality and Social Psychology Review*, 26(2):93–111, May 2022. Publisher: SAGE Publications Inc.
- [177] Kathleen A. Kennedy and Emily Pronin. When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, 34(6):833–848, 2008.
- [178] Lee Ross and Andrew Ward. Naive realism in everyday life: Implications for social conflict and misunderstanding. In *Values and knowledge*, pages 103–135. Psychology Press, 2013.
- [179] Robert J Robinson, Dacher Keltner, Andrew Ward, and Lee Ross. Actual versus assumed differences in construal: “naive realism” in intergroup perception and conflict. *Journal of personality and social psychology*, 68(3):404, 1995.
- [180] Nicholas Epley and Eugene M. Caruso. Perspective taking: Misstepping into others’ shoes. In *Handbook of imagination and mental simulation*, pages 295–309. Psychology Press, New York, NY, US, 2009.
- [181] Isabella Ferreira, Jinghui Cheng, and Bram Adams. The “shut the f** k up” phenomenon: Characterizing incivility in open source code review discussions. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, 2021.
- [182] ShunYi Yeo, Zhuoqun Jiang, Anthony Tang, and Simon Tangi Perrault. Enhancing deliberativeness: Evaluating the impact of multimodal reflection nudges. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA, 2025. Association for Computing Machinery.
- [183] Weiyu Zhang, Tian Yang, and Simon Tangi Perrault. Nudge for reflection: more than just a channel to political knowledge. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2021.
- [184] ShunYi Yeo, Gionnieve Lim, Jie Gao, Weiyu Zhang, and Simon Tangi Perrault. Help me reflect: Leveraging self-reflection interface nudges to enhance deliberativeness on online deliberation platforms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA, 2024. Association for Computing Machinery.

- [185] Elijah L. Claggett, Robert E. Kraut, and Hirokazu Shirado. Relational ai: Facilitating intergroup cooperation with socially aware conversational support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.
- [186] Gauri Kambhatla, Matthew Lease, and Ashwin Rajadesingan. Promoting constructive deliberation: Reframing for receptiveness. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5110–5132, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [187] Jonathan Haidt. The righteous mind: Why good people are divided by politics and religion. *New York Pantheon*, 2012.
- [188] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.
- [189] Pontus Strimling, Irina Vartanova, Fredrik Jansson, and Kimmo Eriksson. The connection between moral positions and moral arguments drives opinion change. *Nature Human Behaviour*, 3(9):922–930, September 2019.
- [190] Michelle Low and Ma. Glenda Lopez Wui. Moral Foundations and Attitudes Towards the Poor. *Current Psychology*, 35(4):650–656, December 2016.
- [191] Dominic Stolerman and David Lagnado. The Moral Foundations of Human Rights Attitudes. *Political Psychology*, 41(3):439–459, 2020.
- [192] Scott Clifford and Jennifer Jerit. How Words Do the Work of Politics: Moral Foundations Theory and the Debate over Stem Cell Research. *The Journal of Politics*, 75(3):659–671, July 2013. Publisher: The University of Chicago Press.
- [193] Marie Altgilbers Roweton and Scott Eidelman. Ideology, moral reframing, and persuasion in the context of COVID-19 vaccines. *Social and Personality Psychology Compass*, 17(9):e12802, 2023.
- [194] Matthew Feinberg and Robb Willer. From Gulf to Bridge: When Do Moral Arguments Facilitate Political Influence? *Personality and Social Psychology Bulletin*, 41(12):1665–1681, December 2015.
- [195] Isaac Baca. Moral Reframing: A Perspective-Taking Approach to Political Arguments. *Scientia et Humanitas*, 11:39–58, June 2021.
- [196] Jonas T. Kaplan, Anthony Vaccaro, Max Henning, and Leonardo Christov-Moore. Moral reframing of messages about mask-wearing during the COVID-19 pandemic. *Scientific Reports*, 13(1):10140, June 2023.
- [197] Joshua L. Kalla, Adam Seth Levine, and David E. Broockman. Personalizing Moral Reframing in Interpersonal Conversation: A Field Experiment. *The Journal of Politics*, 84(2):1239–1243, April 2022.
- [198] Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri. MoralBERT: A Fine-Tuned Language Model for Capturing Moral Values in Social Discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 433–442, September 2024. arXiv:2403.07678 [cs].

- [199] Negar Mokherian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. Moral Framing and Ideological Bias of News. In Samin Aref, Kalina Bontcheva, Marco Braghieri, Frank Dignum, Fosca Giannotti, Francesco Grisolia, and Dino Pedreschi, editors, *Social Informatics*, pages 206–219, Cham, 2020. Springer International Publishing.
- [200] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184, 2020.
- [201] Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. The moral debater: A study on the computational generation of morally framed arguments. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [202] Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. The Moral Foundations Reddit Corpus, August 2022. arXiv:2208.05545 [cs].
- [203] Aaron Isiminger and Roger Giner-Sorolla. Willingness to use moral reframing: Support comes from perceived effectiveness, opposition comes from integrity concerns. *Social Psychological Bulletin*, March 2024.
- [204] Jonathan Haidt and Craig Joseph. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66, 2004.
- [205] Mohammad Atari, Jesse Graham, and Morteza Dehghani. Foundations of morality in iran. *Evolution and Human behavior*, 41(5):367–384, 2020.
- [206] Jesse Graham, Jonathan Haidt, Matt Motyl, Peter Meindl, Carol Iskiwitch, and Marlon Mooijman. Moral foundations theory: On the advantages of moral pluralism over moral monism. 2018.
- [207] Yoo Ji Suh and Hyun Suk Kim. How moral reframing enhances political persuasion: The role of processing fluency and self-affirmation. *Communication Research*, 52(1):61–88, 2025.
- [208] Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. Mapping the Moral Domain. *Journal of personality and social psychology*, 101(2):366–385, August 2011.
- [209] Julia A. Minson, Frances S. Chen, and Catherine H. Tinsley. Why won’t you listen to me? Measuring receptiveness to opposing views. *Management Science*, 66(7):3069–3094, 2020. Place: US Publisher: Institute for Operations Research & the Management Sciences (INFORMS).
- [210] Amanda Baughan, Justin Petelka, Catherine Jaekyung Yoo, Jack Lo, Shiyue Wang, Amulya Paramasivam, Ashley Zhou, and Alexis Hiniker. Someone Is Wrong on the Internet: Having Hard Conversations in Online Spaces. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):156:1–156:22, 2021.
- [211] Lincoln Dahlberg. Rethinking the fragmentation of the cyberpublic: from consensus to contestation. *New Media & Society*, 9(5):827–847, October 2007.

- [212] Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. Studying Moral-based Differences in the Framing of Political Tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:1085–1089, May 2021.
- [213] Ole F. Norheim, Joelle M. Abi-Rached, Liam Kofi Bright, Kristine Bærøe, Octávio L. M. Ferraz, Siri Gloppen, and Alex Voorhoeve. Difficult trade-offs in response to COVID-19: the case for open and inclusive decision making. *Nature Medicine*, 27(1):10–13, January 2021.
- [214] Joe Hoover, Gwennyth Portillo-Wightman, Leigh Yeh, Shreya Havaladar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.
- [215] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346, 2024.
- [216] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc., 2024.
- [217] Jesse Graham, Jonathan Haidt, and Brian A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046, May 2009.
- [218] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. Reddit Rules! Characterizing an Ecosystem of Governance. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), June 2018.
- [219] Charles Kiene, Sohyeon Hwang, Nathan TeBlunthuis, Carl Colglazier, Aaron Shaw, and Benjamin Mako Hill. The Relational Origins of Rules in Online Communities, October 2025. arXiv:2505.18318 [cs].
- [220] Tanvi Bajpai and Eshwar Chandrasekharan. “Think about it like you’re a firefighter”’: Understanding How Reddit Moderators Use the Modqueue. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI ’26, pages 1–16, New York, NY, USA, April 2026. Association for Computing Machinery.
- [221] Sarah A. Gilbert. “I run the world’s largest historical outreach project and it’s on a cesspool of a website.” Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):19:1–19:27, 2020.
- [222] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The Internet’s Hidden Rules: An Empirical Study of

- Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32:1–32:25, November 2018.
- [223] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, July 2019.
 - [224] Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Andreas Damianou, Henrik Lindström, Dan Taber, and Mounia Lalmas. Policy-as-Prompt: Rethinking Content Moderation in the Age of Large Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, pages 840–854, New York, NY, USA, 2025. Association for Computing Machinery.
 - [225] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. LLM-Mod: Can Large Language Models Assist Content Moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA ’24, pages 1–8, New York, NY, USA, 2024. Association for Computing Machinery.
 - [226] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction*, 26(5):1–35, October 2019.
 - [227] MEE6. The Best Discord Bot for Your Server | MEE6. Available from: <https://mee6.xyz/en/> (Accessed: June 17th, 2026).
 - [228] Andrea Wiesner, Svenja Schäfer, and Sophie Lecheler. Navigating the gray areas of content moderation: Professional moderators’ perspectives on uncivil user comments and the role of (AI-based) technological tools. *New Media & Society*, 27(3):1215–1234, March 2025.
 - [229] Shayan Alipour, Shruti Phadke, Seyed Shahabeddin Mousavi, Amirhossein Afsharrad, Morteza Zihayat, and Mattia Samory. The Gray Area: Characterizing Moderator Disagreement on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20 of *ICWSM ’26*, pages 58–75, May 2026.
 - [230] Ali Alkhatib and Michael Bernstein. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.
 - [231] Amanda L. L. Cullen and Sanjay R. Kairam. Practicing Moderation: Community Moderation as Reflective Practice. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–32, March 2022.
 - [232] Agam Goyal, Xianyang Zhan, Yilun Chen, Koustuv Saha, and Eshwar Chandrasekharan. Mo-MoE: Mixture of moderation experts framework for AI-assisted online governance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12645–12660, Suzhou, China, November 2025. Association for Computational Linguistics.
 - [233] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Cross-mod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):174:1–174:30, 2019.

- [234] Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. SLM-Mod: Small Language Models Surpass LLMs at Content Moderation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8774–8790, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.
- [235] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. Watch Your Language: Investigating Content Moderation with Large Language Models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878, May 2024.
- [236] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why johnny can’t prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, New York, NY, USA, 2023. Association for Computing Machinery.
- [237] Youngwoo Kim, Himanshu Beniwal, Steven L. Johnson, and Thomas Hartvigsen. Decoding the rule book: Extracting hidden moderation criteria from Reddit communities. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20487–20498, Suzhou, China, November 2025. Association for Computational Linguistics.
- [238] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):150:1–150:27, 2019.
- [239] Tom Tyler, Matt Katsaros, Tracey Meares, and Sudhir Venkatesh. Social media governance: can social media companies motivate voluntary rule following behavior among their users? *Journal of Experimental Criminology*, 17(1):109–127, March 2021.
- [240] Travis Lloyd, Jennah Gosciak, Tung Nguyen, and Mor Naaman. AI Rules? Characterizing Reddit Community Policies Towards AI-Generated Content. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, pages 1–19, New York, NY, USA, 2025. Association for Computing Machinery.
- [241] Internet Archive: Digital Library of Free & Borrowable Texts, Movies, Music & Wayback Machine.
- [242] Vinay Koshy, Frederick Choi, Yi-Shyuan Chiang, Hari Sundaram, Eshwar Chandrasekharan, and Karrie Karahalios. Venire: A machine learning-guided panel review system for community content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 9(7), October 2025.
- [243] Sara Kiesler, Robert E. Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. In *Building Successful Online Communities*. MIT Press, 2012.
- [244] J. Nathan Matias. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20):9785–9789, May 2019.
- [245] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. Through the Looking Glass: Study of Transparency in Reddit’s Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–35, January 2020.

- [246] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction*, 30(1):1–34, February 2023.
- [247] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4. McLean, VA, USA, 2005.
- [248] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, September 1996. ISSN: 1049-2615.
- [249] Zana Buşinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):188:1–188:21, 2021.
- [250] danah boyd. You Think You Want Media Literacy... Do You?, March 2018.
- [251] Crystal Lee, Tanya Yang, Gabrielle D Inchoco, Graham M. Jones, and Arvind Satyanarayan. Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, pages 1–18, New York, NY, USA, May 2021. Association for Computing Machinery.
- [252] Farnaz Jahanbakhsh and David R Karger. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, pages 1–21, New York, NY, USA, 2024. Association for Computing Machinery.
- [253] Arlen Fan, Fan Lei, Steven R. Corman, and Ross Maciejewski. Skeptik: A Hybrid Framework for Combating Potential Misinformation in Journalism. *ACM Transactions on Interactive Intelligent Systems*, 16(1):1:1–1:31, 2026.
- [254] Jane Mansbridge, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F. Thompson, and Mark E. Warren. A systemic approach to deliberative democracy. In Jane Mansbridge and John Parkinson, editors, *Deliberative Systems: Deliberative Democracy at the Large Scale*, Theories of Institutional Design, pages 1–26. Cambridge University Press, Cambridge, 2012.
- [255] Jane Mansbridge, James Bohman, Simone Chambers, David Estlund, Andreas Føllesdal, Archon Fung, Cristina Lafont, Bernard Manin, and José luis Martí. The place of self-interest and the role of power in deliberative democracy. *Journal of Political Philosophy*, 18(1):64–100, 2010.
- [256] Lynn M. Sanders. Against deliberation. *Political theory*, 25(3):347–376, 1997.
- [257] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(Volume 22, 2019):129–146, May 2019.
- [258] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. Affect, Not Ideology. *Public Opinion Quarterly*, 76(3):405–431, September 2012.

- [259] James N Druckman and Matthew S Levendusky. What Do We Measure When We Measure Affective Polarization? *Public Opinion Quarterly*, 83(1):114–122, May 2019.
- [260] Matthew Feinberg and Robb Willer. The Moral Roots of Environmental Attitudes. *Psychological Science*, 24(1):56–62, January 2013.
- [261] Lala Muradova. Seeing the Other Side? Perspective-Taking and Reflective Political Judgements in Interpersonal Deliberation. *Political Studies*, 69(3):644–664, August 2021. Publisher: SAGE Publications Ltd.
- [262] D. A. Prentice and D. T. Miller. Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm. *Journal of personality and social psychology*, 64(2):243–256, February 1993.
- [263] Diana C. Mutz. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press, 2006.
- [264] James S. Fishkin. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford University Press, Oxford ; New York, 2009.

Acknowledgment

I was able to reach the end of this long journey only with the understanding, support, and care of so many people around me. I specifically want to express my gratitude to:

- **My advisor, Juho Kim.** Thank you so much for all the support, advice, guidance, inspiration, and patience throughout this journey. You not only taught me to be an independent researcher but also inspired me to be a better person with your kind and open attitude to everyone, your commitment to real-world impact, and your boldness in taking risks and pursuing ambitious ideas. It was truly fortunate for me to work with and learn from you.
- **My amazing thesis committee, Jeong-woo Jang, Brian McInnis, Joseph Seering, and Amy X. Zhang.** I deeply appreciate your thoughtful feedback, advice, and encouragement to improve my dissertation. Each of you brought a distinct lens that pushed me to articulate my work more coherently and think more carefully about its broader implications.
- **All of my collaborators,** including **Haesoo, Duy, Gionnieve, Yoojin, Vedad, and Eunhye.** Your creativity, constructive feedback, and countless hours spent building systems and running studies made this work possible. Thank you for sharing both the excitement of research and the moments we got lost in it together. I owe much to **Juhoon** for her sharp and thoughtful feedback, for helping recruit participants, and for bringing deep knowledge of Reddit and online moderation that made the DecideTogether project so much richer. I also thank my senior collaborators, **Prof. Dae Hyun Kim, Prof. Elisabeth Lex, Prof. Minsu Park, Dr. Jin Young Kim, Prof. Meng Xia,** and **Prof. Jihee Kim,** for their thoughtful advice and guidance that have broadened my perspective on research.
- **All past and present KIXLAB members.** Being part of KIXLAB has been the luckiest part of this journey—it was a place where I always felt supported, free to fail, and pushed to grow. **Yoonseo,** you were all of these at once: the most skillful collaborator, the most thoughtful reviewer, the most caring supporter, and the best friend I could have asked for in the lab. I owe so much of this journey to your wisdom, your encouragement, and the time we shared. **Eun-Young,** I am grateful to have you as my first and closest collaborator and as a senior who shares my research interests. Navigating our first project together shaped how I think about research. Your warm welcome on my first day at KIXLAB—which happened to be your first day as well—is something I will never forget. **Kihoon,** the *ahjussi* restaurants and climbing gyms were exactly what I needed. I’m glad you kept dragging me out. **Taesoo,** I lost count of how many times your research and technical advice saved me. I don’t know how I would have gotten through without you. **Yoonjoo** and **Saelyne,** watching you consistently find and pursue impactful research questions was quietly inspiring. The dinners out, the invitations home, and the late-night chats meant a lot to me. **Yoonsu,** your energy was impossible to contain and somehow contagious in the best way. It made the lab feel much friendlier and more fun. **DaEun** and **Hyehyun,** our lunch breaks full of laughter from random stories and the latest MZ trends were some of the best hours in the lab. **Bekzat,** the endless research discussions and the random aggressive jokes were among the moments I enjoyed most in the lab. **Yeon Su,** I owe you so much for stepping up and taking on the grant work

during my final stretch. **SIG-baseball** members, including **Heechan**, **Junho**, and **Jaesang**, I am grateful for sharing the joy and pain that only baseball can bring. **Hyungyu**, **Seoyoung**, and **Minsuk**, you set the lab's tone and did much of the heavy lifting early on, and I learned so much about doing research from each of you. And **Ms. Oaksil Kim**, I appreciate how you navigated every impossible administrative request with grace.

- **My dearest friends.** Having you all around gave me a life outside of this journey, and the support to keep going within it. **Youngjae**, thank you for showing me whole new worlds of alcohol and cinema, for sharing the weekend meals and late-night chats, and for enduring my messiness for eight years. This doctoral journey felt much less lonely because of you. **Dongsu** and **Sunghyun**, every time I showed up in Seoul out of nowhere, you made time for me. Those were some of the best chances I had to breathe. **Seungho**, **Yooju**, and **Eui**, the meals and the time we carved out together meant more to me than you know.
- **The small things that kept a little room open to breathe.** My Pikmin, staring blankly back at me, kept me walking, in both senses. The coffee machine stood by through the early mornings and the late nights. And the Lotte Giants taught me that every season ends eventually, no matter how badly you play through it.
- **The citizens of Korea** who stood with candles and lightsticks through the darkest and coldest nights of 2016 and 2024. Everyone's courage and determination to make the world a better place motivated me to start this research.
- **My family.** I thank my parents, **Keunhee** and **Myungyong**, for their unconditional love, support, and belief in me. I was able to overcome all those challenges and turbulent times because of your support. 특별히 우리 할머니 **김범순** 여사님, 할머니의 끝없는 사랑과 수없는 기도 덕분에 할머니 회사에 따라다니던 아이가 커서 여기까지 올 수 있었어요. 항상 감사합니다. 제가 잘 할게요! I also thank my uncle **Sooyoung** for his unwavering support in every sense. **Siyoun**, you were always in my corner, through the trips, the pointless conversations, and everything in between.