Deliberative Machines: From Reflective Dialogue to Fair Consensus with Language Models and Social Choice

by

Carter Blair

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2025
© Carter Blair 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Carter Blair was the sole author of Chapter 1 and Chapter 4 which were written under the supervision of Dr. Kate Larson and Dr. Edith Law.

Additionally, this thesis consists in part of two manuscripts written for publication that are currently under review. Exceptions to sole authorship of material are as follows:

Chapter 2: The research in this chapter was supervised by Dr. Kate Larson and Dr. Edith Law.

Chapter 3: The research in this chapter was supervised by Dr. Kate Larson.

Abstract

This thesis investigates the bidirectional relationship between artificial intelligence (AI), particularly large language models (LLMs), and social choice theory. Firstly, it explores how principles from social choice can address challenges in AI alignment, specifically the problem of aggregating diverse human preferences fairly when guiding AI behavior (**SC** $\rightarrow$ **AI**). Standard alignment methods often obscure value conflicts through implicit aggregation. Secondly, it examines how AI techniques can enhance collective decision-making processes traditionally studied in social choice (**AI** $\rightarrow$ **SC**), offering new ways to elicit and synthesize the complex, nuanced, and verbal preferences that conventional mechanisms struggle to handle.

To address these issues, this work presents computational methods operating at the interface of AI and social choice. First, it introduces Interactive-Reflective Dialogue Alignment (IRDA), a system using LLMs to guide users through reflective dialogues for preference elicitation. This process helps users construct and articulate their values concerning AI behavior, resulting in individualized reward models that capture preference diversity with improved accuracy and sample efficiency compared to non-reflective baselines, especially when values are heterogeneous.

Second, the thesis proposes a framework for generating fair consensus statements from multiple viewpoints by modeling text generation as a token-level Markov Decision Process (MDP). Within this MDP, agent preferences are represented by policies derived from their opinions. We develop mechanisms grounded in social choice: a stochastic policy maximizing proportional fairness (Nash Welfare) to achieve ex-ante fairness guarantees (1-core membership) for distributions over statements, and deterministic search algorithms (finite lookahead, beam search) maximizing egalitarian welfare for generating single statements. Experiments demonstrate that these search methods produce consensus statements with better worst-case agent alignment (lower Egalitarian Perplexity) than baseline approaches.

Together, these contributions offer principled methods for eliciting diverse, reflective preferences and synthesizing them into collective outputs fairly. The research provides tools and insights for developing AI systems and AI-assisted processes that are more sensitive to value pluralism.

Acknowledgements

I am deeply grateful to my supervisors, Kate Larson and Edith Law, for their guidance throughout this research. I am thankful for their consistent willingness to engage with nascent ideas and for their skill in helping me identify the core themes that ultimately provided the thesis with its direction and coherence. Their guidance was also characterized by a remarkable ability to foresee connections and potential directions; I frequently found myself 'discovering' something, only to realize it stemmed from a suggestion they had made months prior. Underlying their technical advice was a clear focus on the research's real-world applicability and potential positive impact. This commitment to developing technology that can improve people's lives is a goal that has inspired me and that I hope to always work towards.

I thank my colleagues in the Multi-Agent Systems Group. I am particularly grateful to Ben for his help and guidance during my first project at Waterloo and throughout my degree. Nanda was a great lab neighbor and valuable conversation partner, and I appreciate the many times he helped me work through research problems. To the rest of the lab members, thank you for the camaraderie and the great discussions during our meetings.

My thanks also go to the members of the HCI Lab. Their perspectives were particularly helpful in designing and interpreting the user studies in Chapter 2. I benefited greatly from the friendship and research discussions within the lab, including with Ryan, Nikhita, Ludwig, Linda, and Matthew.

I thank my committee members, Robin Cohen and Yuntain Deng, for their time and effort in reading this thesis.

I owe a significant debt of gratitude (which I will never be able to adequately repay) to my parents and grandparents for their love and support (emotional, practical, and financial). As well as for always encouraging me to further my education and pursue what I love.

Lastly, I am especially grateful to my beautiful fiancée Kate for the immense amount of dopamine I have received from her, which has been my main source of fuel throughout my master's. More seriously, I would like to thank her for fleshing out ideas with me, unquestioningly picking up the real-life slack (especially before paper deadlines), and her constant encouragement, hype, and love.

Dedication

I would like to dedicate this thesis to my wonderful parents for all of their love, unconditional support of my winding academic path, and for the countless other ways in which they have supported me and my education over the years.

Table of Contents

List of Figures x

List of Tables xi

1 Introduction 1
- 1.1 Thesis Statement . . . 2
- 1.2 Contributions . . . 2
- 1.3 Thesis Overview . . . 3

2 Reflective Reward Design for Pluralistic Alignment 4
- 2.1 Motivation . . . 4
- 2.2 Related Work . . . 6
- 2.3 Method . . . 7
- 2.4 Study Design . . . 10
  - 2.4.1 Environments . . . 10
  - 2.4.2 Participants . . . 11
  - 2.4.3 Procedure . . . 11
  - 2.4.4 Baseline Comparisons . . . 12
  - 2.4.5 Analysis . . . 13
- 2.5 Results: Study 1 - Multi-Agent Apple Farming . . . 15
- 2.6 Results: Study 2 - The Moral Machine . . . 17
- 2.7 Discussion . . . 20

3 Consensus Statement Generation via Token-Level Tree Search 22
- 3.1 Motivation . . . 22
- 3.2 Related Work . . . 23
- 3.3 Problem Setup & Preliminaries . . . 25
- 3.4 Stochastic Policies and Ex-Ante Fairness . . . 28
  - 3.4.1 Computational Tractability via Token Chunking . . . 29
  - 3.4.2 Deriving the Ex-Ante Fair Stochastic Policy . . . 29
  - 3.4.3 Properties of the Induced Policy . . . 30
  - 3.4.4 A Note on Computation and Approximation . . . 32
- 3.5 Generating a Single Consensus Statement . . . 32
  - 3.5.1 The Rawlsian Path . . . 33
- 3.6 Experiments . . . 35
  - 3.6.1 Evaluating Prompt-Based Credit Assignment . . . 35
  - 3.6.2 Consensus Generation . . . 36
- 3.7 Discussion . . . 38

4 Conclusion & Future Work 40
- 4.1 Future Work . . . 41
  - 4.1.1 Future Work: AI for Social Choice (AI → SC) . . . 41
  - 4.1.2 Future Work: Social Choice for AI (SC → AI) . . . 42
- 4.2 Concluding Remarks . . . 43

Bibliography 44

Appendices **49**

A Appendix to Chapter 2 **50**
A.1 Prompts and ASCII Encodings 50
A.1.1 Study 1: Multi-Agent Apple Farming Environment 50
A.1.2 Study 2: The Moral Machine 53
A.2 Descriptions of Participant Decision Making Features 55
A.2.1 Study 1: Multi-Agent Apple Farming Environment 55
A.2.2 Study 2: The Moral Machine 56

B Appendix to Chapter 3 **57**
B.1 Prompts 57
B.2 Credit Assignment Results for Gemma 58
B.3 Scenarios and Sample Consensus Statements 59

List of Figures

1.1 Bidirectional relationship between Social Choice Theory and AI techniques studied in this thesis. 1

2.1 Overview of the RLHF pipeline with *Interactive-Reflective Dialogue Alignment*. A more detailed description is provided in Algorithm 1. 7
2.2 ASCII encoding of two timesteps of a trajectory of the multi-agent apple farming environment. 8
2.5 Study procedure for both studies in Chapter 2. 11
2.3 Participant familiarity with reinforcement learning in Study 1 of Chapter 2. 11
2.4 Participant familiarity with reinforcement learning in Study 2 of Chapter 2. 11
2.6 CNN architecture used in Study 2 of Chapter 2: Feature Extraction (Conv/Pool layers) feeds into Classification (FC layers). Output neurons for 'Stay' and 'Swerve.' 12
2.7 (Left) Distributions of balanced accuracies for language-based reward models: our pipeline (IRDA) vs. baseline (L^B) in Study 1 of Chapter 2. (Right) Distribution of per-participant differences in balanced accuracy ($P_i^{\text{IRDA}} - P_i^B$) between IRDA and baseline models in Study 1 of Chapter 2. . . 14
2.8 Balanced accuracy of models vs. samples per participant in **Study 1** of Chapter 2. Blue line shows average individual MLP (MLP^{ind}); red shows collective MLP (MLP^{col}). Our IRDA system (green dashed) and baseline (L^B, yellow dashed) used 4 samples per participant (vertical gray line). Collective model trained on 21x samples shown (21 participants). Shaded areas: 95% confidence intervals. Gray dashed line: random performance. 15
2.9 (Left) Distributions of accuracies for language-based reward models: our pipeline (IRDA) vs. baseline in Study 2. (Right) Distribution of per-participant differences in accuracy ($P_i^{\text{IRDA}} - P_i^B$) between IRDA and baseline models in Study 2. 18
2.10 Comparison of model accuracies as a function of samples per participant in Study 2. (Left) Performance of MLP-based models: average individual MLP (MLP^{ind}, blue), collective MLP (MLP^{col}, red), our IRDA approach (green dashed), and baseline L^B (yellow dashed). (Right) Performance of CNN-based models: average individual CNN (CNN^{ind}, blue), collective CNN (CNN^{col}, red), IRDA, and baseline. Both panels show confidence intervals (shaded areas), the 6-sample training point for IRDA and baseline (vertical gray line), and the random performance level (gray dashed). 19

3.1 Illustration of the token-level generation tree. Each state s_t is a partial sequence. Edges represent actions (chosen tokens). Paths ending in a boxed node represent completed sequences ending with $\langle\text{eos}\rangle$, which belong to the set $\mathscr{C}$. Paths ending in vertical dots represent incomplete branches. . 26
3.2 Illustration of the induced stochastic policy Π^* at state s. The optimal lottery p^* assigns probabilities to the leaf nodes (complete paths). The probability $P^*(\mathscr{C}_{\mathscr{K}}(s))$ is the sum of $p^*(X)$ for all leaves reachable from s. The probability $P^*(\mathscr{C}_{\mathscr{K}}(s,k))$ is the sum for leaves reachable via action k. The policy $\Pi^*(s,k)$ is the conditional probability of taking action k. 30
3.3 Per-scenario egalitarian welfare (perplexity). Lower values indicate better minimum agent utility. Striped bars indicate that the method uses search over the token-level MDP. Results are also shown in Table 3.2. 38

A.1 Prompt used to generate the hypothesis and alternative perspectives in Study 1 of Chapter 2. . . 51
A.2 Prompt used to generate the trajectory label (reward) in Study 1 of Chapter 2. 52
A.3 ASCII encoding of a Moral Machine scenario from Study 2. 53
A.4 Prompt used to generate the hypothesis and alternative perspectives in Study 2 of Chapter 2. . 54
A.5 Prompt used to generate the trajectory label (reward) in Study 2 of Chapter 2. 55

B.1 Reference policy prompt used in Chapter 3. 57

B.2 Agent policy prompt used in Chapter 3. 57
B.3 Brush up prompt used in Chapter 3. 58

List of Tables

2.1 Mapping of Analysis Methods to Research Questions 13
2.2 Behavioural features participants used to decide whether the agent was acting respectfully in Study 1 of Chapter 2. Cells coloured in light grey ( ) indicate that the participant did not consider the feature. Cells colored in green ( ) indicate the participant considered the feature in their decision making. 16
2.3 Behavioural features participants considered (Study 2). Cells coloured in light grey ( ) indicate that the participant did not consider the feature. Cells colored in green ( ) indicate the participant considered the feature. 18

3.1 Credit assignment results for `Llama 3.1 8B Instruction-Tuned`. Darker green indicates larger Z-score. Z-score column is for altered tokens. Alterations are represented by "<misaligned>/<aligned>". 35
3.2 Egalitarian Welfare (Perplexity) Across All Scenarios. Lower is better. 39

B.1 Credit assignment results for `Gemma 2 9b Instruction-Tuned`. Darker green indicates larger Z-score. Z-score column is for altered tokens. Alterations are represented by "<misaligned>/<aligned>". 59
B.2 Summary of Agent Opinions and Method Consensus on Scenario 1: "*Are we living in a golden age of democracy?*" 60
B.3 Summary of Agent Opinions and Method Consensus on Scenario 2: "*Should the UK's ties to Europe be stronger or weaker?*" 61
B.4 Summary of Agent Opinions and Method Consensus on Scenario 3: "*Is it okay to use animals for food?*" 62

It is not from ourselves that we will learn to be better than we are.

— Wendell Berry, *A Native Hill* (1968)

Introduction 1

1.1 Thesis Statement 2
1.2 Contributions 2
1.3 Thesis Overview 3

The aggregation of individual preferences into collective decisions is a central problem studied within social choice (SC) theory [Bra+16]. This theoretical foundation becomes increasingly relevant with the development of Artificial Intelligence (AI) and the associated non-trivial problem of aligning AI behaviour with human preferences. Specifically, the task of aligning a single AI system with the preferences of multiple human principals raises core social choice questions [Ge+24; Con+24]. It requires navigating value pluralism and determining how to synthesize the diverse inputs of many principals into a coherent objective to guide an AI's behaviour. While AI development poses new challenges for social choice theory, AI techniques also offer novel capabilities for handling preferences expressed in open-ended, verbal formats, which have traditionally been difficult to elicit and aggregate systematically [Fis+24]. This thesis investigates the relationship between social choice and AI, focusing on two primary directions: the application of social choice principles to AI alignment and the use of AI techniques to improve collective decision-making mechanisms (Figure 1.1).

[Ge+24]: Ge et al. (2024), 'Axioms for AI Alignment from Human Feedback'
[Con+24]: Conitzer et al. (2024), 'Position: social choice should guide AI alignment in dealing with diverse human feedback'

[Fis+24]: Fish et al. (2024), 'Generative Social Choice'

Focusing on the first direction (**SC → AI**), insights from social choice theory are directly applicable to the problem of pluralistic AI alignment. Ensuring that AI systems act in accordance with the diverse values of multiple principals necessitates methods for fair preference aggregation, a core concern of social choice. One approach, which is developed in this work, involves creating personalized or representative preference/reward models that capture the nuances of individual viewpoints. Such explicit models form a foundation for applying transparent, principled aggregation rules to determine the AI's collective objective. This contrasts with methods where aggregation is implicit or opaque, and moves towards a 'virtual democracy' [Noo+18] model for alignment, where the AI's goals are derived systematically from well-modeled constituent preferences.

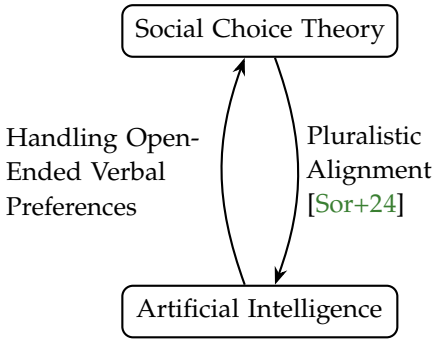


Figure 1.1: Bidirectional relationship between Social Choice Theory and AI techniques studied in this thesis.

[Noo+18]: Noothigattu et al. (2018), 'A voting-based system for ethical decision making'

Turning to the second direction (**AI → SC**), AI methods, particularly large language models (LLMs), offer novel capabilities that can improve collective decision-making processes and our ability to "measure collective will." This potential contrasts sharply with the limitations inherent in traditional tools like opinion polls or ballots. Opinion polls and ballots often rely on structured formats that struggle to capture nuanced, open-ended opinions. Their design can also introduce framing effects [BM07] and obscure the collective will by constraining choice to predefined options. This power to define the available choices—the power to set the agenda—is crucial yet often overlooked, as elaborated in the following passage from Landemore:

[BM07]: Bütler et al. (2007), *Framing effects in political decision making: Evidence from a natural voting experiment*

> *"Agenda-setting power—or 'non-decision-making power,' as it is also called—is the less visible 'second face' of power so compellingly painted by Peter Bachrach and Morton Baratz (1962; see also Lukes 1974). Agenda-setting power is as essential to power, if not more so, as casting the final vote—the first, most visible face of power, which most people focus on.* ***Given that issues or options to be voted on cannot formulate themselves, someone needs to be in charge of doing just that and,***

> ***in the process, is bound to exercise an enormous amount of power.*** *[Emphasis added]" [Lan20, p. 60]*

[Lan20, p. 60]: Landemore (2020), *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*

Furthermore, opinions gathered via traditional polling or through static ballots may lack depth, representing "top-of-the-head" responses rather than considered judgments [FL05]. Research on deliberative methods, such as Deliberative Polling, demonstrates that when provided with balanced information and structured opportunities for discussion and reflection, individuals can develop more informed and stable preferences, distinct from those captured by less interactive elicitation methods [FL05]. LLMs present an avenue for facilitating similar processes; their capacity for natural language interaction allows them to aid in more interactive preference elicitation, support the reflective processes needed for individuals to articulate underlying values, and present alternative perspectives or relevant factual information. However, eliciting such nuanced preferences is only valuable when accompanied by corresponding methods to process the resulting free-form text. LLMs contribute here as well, offering capabilities that, when aided by tools from social choice theory, can synthesize such input into a collective output[1]. Nevertheless, it is important to recognize that LLMs inherit biases from their training data and may possess emergent political leanings [Fen+23]. Furthermore, their operational complexity can hinder interpretability and accountability, posing risks if their outputs are treated as definitive representations of collective will rather than as potentially useful, but fallible, tools for collective sense-making [RP25].

[FL05]: Fishkin et al. (2005), 'Experimenting with a democratic ideal: Deliberative polling and public opinion'

1: Our method for doing so is presented in Chapter 3.

[Fen+23]: Feng et al. (2023), 'From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models'

[RP25]: Revel et al. (2025), 'AI-Facilitated Collective Judgements'

1.1 Thesis Statement

Building upon this potential and addressing the identified limitations of traditional methods, this thesis argues for the following:

Thesis Statement

Using LLMs to connect preference elicitation, diversity-preserving modeling, and nuanced consensus generation offers possibilities for advancing how diverse human preferences are incorporated into collective decision-making (addressing the **AI** → **SC** direction), particularly by eliciting and handling open-ended verbal input; furthermore, these methods, particularly those for preference elicitation and diversity-preserving modeling, are also valuable for addressing challenges in pluralistic AI alignment (providing a foundation for the **SC** → **AI** direction).

1.2 Contributions

Specifically, this research contributes computational methods focusing on: (1) accessible, LLM-driven conversational techniques designed for eliciting rich, reflective verbal preferences, moving beyond restrictive formats (Chapter 2); (2) individualized frameworks, using the few-shot capabilities of LLMs, applied to model the diversity inherent in these preferences, thereby supporting explicit aggregation distinct from

the implicit, potentially homogenizing aggregation characteristic of standard reinforcement learning from human feedback (Chapter 2); and (3) techniques for generating consensus statements by guiding LLM generation with policies representing constituent perspectives, where these policies are derived either directly from verbal preferences or from the individualized reward models (Chapter 3). Overall, these contributions provide computational tools and insights towards AI systems and AI-assisted processes designed to be sensitive to value pluralism while allowing individuals greater flexibility in expressing their unconstrained perspectives.

1.3 Thesis Overview

The remainder of this thesis proceeds as follows: Chapter 2 details the first core project on Reflective Reward Design for Pluralistic Alignment, presenting the IRDA system for eliciting individualized preferences via reflective dialogue and its empirical evaluation. Chapter 3 describes the second core project on fair consensus statement generation, outlining the MDP-based framework, proportional fairness criteria, and associated algorithms for generating consensus text from multiple policies as well as theoretical results and an empirical evaluation of various consensus generation methods. Chapter 4 offers a synthesis and discussion, analyzing the connections between the two projects and their combined implications for addressing value pluralism in AI. The chapter concludes by summarizing the contributions, acknowledging limitations, and suggesting directions for future research. A bibliography and appendices follow.

Reflective Reward Design for Pluralistic Alignment

2

Agents, whether robots or language models, are commonly aligned with "human values" through reinforcement learning from human feedback (RLHF), where a single reward model is learned from aggregated human feedback and used to align an agent's behavior. However, human values are not homogeneous–different people hold distinct and sometimes conflicting values. Aggregating feedback into a single reward model risks disproportionately suppressing minority preferences and unique perspectives. To address this, we present a novel reward modeling approach for learning *individualized* reward models. Our approach uses a language model to guide users through *reflective* dialogues where they critique agent behavior and construct their preferences. This personalized dialogue history, containing the user's reflections and critiqued examples, is then used as context for another language model that serves as an individualized reward function for evaluating new trajectories. In two studies with a total of 30 participants, our method achieved a 9-12% improvement in accuracy over non-reflective language-based reward models while being vastly more sample efficient than traditional supervised learning methods.

2.1 Motivation 4
2.2 Related Work 6
2.3 Method 7
2.4 Study Design 10
2.4.1 Environments 10
2.4.2 Participants 11
2.4.3 Procedure 11
2.4.4 Baseline Comparisons . . 12
2.4.5 Analysis 13
2.5 Results: Study 1 - Multi-Agent Apple Farming . . 15
2.6 Results: Study 2 - The Moral Machine 17
2.7 Discussion 20

2.1 Motivation

As AI systems gain more widespread adoption and use, the need to align them with "human values" has gained urgency. Yet, human values are deeply varied[1]. However, the predominant method for aligning agent, RLHF, typically relies on training a *single* reward model to capture preferences aggregated from multiple users [Chr+17; Ouy+22]. These monolithic reward models implicitly make trade-offs among competing values and can disproportionately suppress minority viewpoints [SLH23; Cha+24].

In response to this shortcoming, recent work has explored approaches that better preserve the diversity of human values [SLH23; Cha+24; Pod+24]. While the specifics differ, these approaches broadly increase the granularity of the reward signal to account for differences in opinion. For example, some methods learn multiple distinct reward models to capture different preference clusters [Cha+24]. Others learn reward models that output distributions rather than scalar values to represent uncertainty and variation in preferences [SLH23], and others learn *individualized* reward models for each person [Pod+24].

These higher-granularity approaches offer benefits for collective and individual alignment. In collective settings where we must aggregate preferences across many stakeholders, having granular representations of preferences enables making transparent choices about balancing competing values by explicitly aggregating rewards. Rather than implicitly averaging preferences, we can optimize for desired properties like egalitarian welfare. Meanwhile, for personal AI assistants, techniques that learn high granularity preference representations can give rewards that

1: Shwartz provides the following example: *'[. . .] students and teachers from the United States are conspicuous for the high importance they attribute to values expressing a desire to get ahead personally in the social hierarchy [. . .] and the low importance they attribute to values expressing social concern [. . .]. Spanish and Italian students and teachers show the opposite pattern.'* [Sch92, footnote 19, p. 60]

[Chr+17]: Christiano et al. (2017), 'Deep reinforcement learning from human preferences'

[Ouy+22]: Ouyang et al. (2022), 'Training language models to follow instructions with human feedback'

[SLH23]: Siththaranjan et al. (2023), 'Understanding Hidden Context in Preference Learning: Consequences for RLHF'

[Cha+24]: Chakraborty et al. (2024), 'MaxMin-RLHF: alignment with diverse human preferences'

[Pod+24]: Poddar et al. (2024), 'Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning'

are more representative of the given user compared to the population average. However, a key challenge exists: While a single reward model can leverage data from all users to learn a scalar reward, higher-granularity approaches must learn more complex patterns (e.g., preference distributions [SLH23] or personalized reward functions [Pod+24]), which are more challenging to estimate due to their higher dimensionality and sparser per-pattern data coverage.

A further complication arises from the process of *preference construction* in complex and novel domains, where humans actively construct their preferences rather than simply reveal them. While current reward modeling approaches typically rely on passive data collection and straightforward labeling, research indicates this may be insufficient: the process of *constructing* preferences benefits significantly from deliberate reflection, which helps individuals transform their latent values into concrete preferential choices [Fis91]. This insight is further supported by findings from psychology and consumer research, which demonstrate that prompting people to actively contemplate their underlying values and reasoning processes leads to more stable and well-defined preferences[2] [HDD14; VVH+20].

To address these two challenges, we introduce *Interactive-Reflective Dialogue Alignment* (IRDA), a system that uses large language models (LLMs) to learn personalized reward functions through interactive dialogue. IRDA combines three core components: (1) **reflective language-based preference elicitation** that guides users in articulating their values, (2) **active learning** to strategically select examples for human critique, and (3) **LLM-driven reward modeling** where the LLM directly serves as the reward function by leveraging its in-context learning capabilities to generalize from sparse user feedback. IRDA's architecture directly confronts the preference construction problem by replacing passive labeling with LLM-guided dialogues that provoke deliberate, context-sensitive reflection. At the same time, its data-efficient learning strategy [3] mitigates the difficulty of learning personalized reward functions.

We evaluated IRDA with two user studies involving 30 participants in total. The first study (21 participants) focused on building a reward model for each user's personal definition of "respectful behavior," while the second (9 participants) explored ethical decision-making in autonomous vehicles. Across both studies, participants differed widely in their value judgments, and our system was able to capture these individual definitions of value-aligned behavior more accurately than baseline approaches.

Our contributions are as follows:

- A novel pipeline for aligning AI agents to individual values, informed by AI, HCI, and social science.
- A comprehensive evaluation in two distinct domains, demonstrating that our system captures individual preferences more accurately than baselines.
- Empirical characterization of how individuals diverge in their conceptions of value-aligned AI behavior.
- Insights for future research on interactive systems that help end users construct, refine, and operationalize their latent values for both personal and collective alignment.

2: One can draw a strong connection between framing effects in voting as described by Bütler and Maréchal [BM07], and unstable, surface-level preferences. Presumably, if voters' preferences were deeply considered and stable, framing effects would not exist.
[HDD14]: Hauser et al. (2014), 'Self-reflection and articulated consumer preferences'
[VVH+20]: Ver Donck et al. (2020), 'Improving patient preference elicitation by applying concepts from the consumer research field: narrative literature review'

3: In particular, through the use of active learning and the few-shot learning capabilities of LLMs

The rest of this chapter looks at the related work (Section 2.2), describes the details of the system (Section 2.3), outlines the experimental procedure (Section 2.4) and the findings of both studies (Section 2.5, Section 2.6), and concludes with a discussion of the results (Section 2.7).

2.2 Related Work

Pluralistic Alignment. Recent work has highlighted the importance of moving beyond monolithic reward models and toward approaches that capture heterogeneous or uncertain human preferences in AI alignment. For instance, distributional preference learning (DPL) estimates an entire distribution over possible reward values, thereby accommodating hidden context and diverse annotator criteria [SLH23]. Similarly, methods like MaxMin-RLHF learn a mixture of reward models and optimize an egalitarian objective to avoid disproportionately favoring majority viewpoints [Cha+24]. Others have proposed user-specific latent variables that personalize reward models without requiring extensive per-user labels [Pod+24] or have leveraged meta-learning to reduce feedback requirements [HS22]. However, these approaches assume users have direct access to their preferences in novel contexts despite evidence from suggesting otherwise [HDD14]. Our work complements these advances by actively eliciting fine-grained, user-specific preferences and helping users turn their latent values into concrete preferences through guided reflection.

[SLH23]: Siththaranjan et al. (2023), 'Understanding Hidden Context in Preference Learning: Consequences for RLHF'

[Cha+24]: Chakraborty et al. (2024), 'MaxMin-RLHF: alignment with diverse human preferences'

[Pod+24]: Poddar et al. (2024), 'Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning'

[HS22]: Hejna et al. (2022), 'Few-Shot Preference Learning for Human-in-the-Loop RL'

[HDD14]: Hauser et al. (2014), 'Self-reflection and articulated consumer preferences'

Language-Based Reward Design. A separate line of research explores using large language models (LLMs) to specify or generate reward functions. Some methods prompt LLMs to propose reward code, which is then used to train RL policies via standard optimization [Ma+24; Xie+24; Ver+24; Beh+24]. Other methods directly treat an LLM as a proxy reward function by prompting it with desired behavior descriptions [Kwo+23]. These LLM-based approaches have made reward specification more accessible, particularly in domains where handcrafting objectives is difficult. For example, as Kwon et al. [Kwo+23] note, specifying a reward function by hand to capture a desired verbal negotiation style, such as being a "versatile negotiator", is challenging, but LLMs can interpret examples or descriptions of such styles. While these methods make specifying a reward function more accessible, they treat it as a one-way street, where users tell the LLM what they want. Our system makes it a two-way dialogue, using LLMs both to help users clarify their preferences and to translate those preferences into reward functions.

[Ma+24]: Ma et al. (2024), 'Eureka: Human-Level Reward Design via Coding Large Language Models'

[Xie+24]: Xie et al. (2024), 'Text2Reward: Reward Shaping with Language Models for Reinforcement Learning'

[Ver+24]: Verma et al. (2024), 'Balancing Act: Prioritization Strategies for LLM-Designed Restless Bandit Rewards'

[Beh+24]: Behari et al. (2024), 'A Decision-Language Model (DLM) for Dynamic Restless Multi-Armed Bandit Tasks in Public Health'

[Kwo+23]: Kwon et al. (2023), 'Reward Design with Language Models'

Reflection as a Path to Expressing Latent Preferences. Preferences are rarely pre-defined artifacts waiting to be extracted; instead, in new contexts, they form through reflective processes that turn latent values into concrete preferences [Fis91]. In consumer and behavioral research, explicitly prompting users to reflect on trade-offs or alternative perspectives fosters more stable and revealing preference statements [HDD14; VVH+20]. We build on this insight by weaving reflection into an LLM-based alignment pipeline, enabling users to clarify and externalize their values for AI systems.

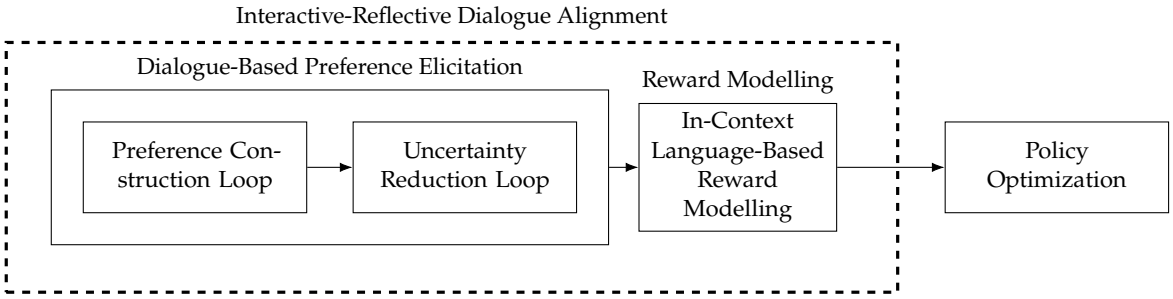


Figure 2.1: Overview of the RLHF pipeline with *Interactive-Reflective Dialogue Alignment*. A more detailed description is provided in Algorithm 1.

Designing Reflection into Dialogue Systems. Within HCI, frameworks like Fleck and Fitzpatrick's [FF10] outline how technologies can scaffold reflective thinking by identifying specific activities (such as revisiting experiences or seeking explanations) that technology can support. Studies also show that structured verbal prompts support deeper introspection over time [Koc+18; Wol+22]. Recent advances in LLM-driven agents further enable flexible interactive reflective dialogues [AY24]. Our contribution is to extend these dialogues to preference elicitation for AI alignment, using LLMs to help users iteratively refine how an AI agent should act.

[FF10]: Fleck et al. (2010), 'Reflecting on reflection: framing a design landscape'

[Koc+18]: Kocielnik et al. (2018), 'Reflection Companion'
[Wol+22]: Wolfbauer et al. (2022), 'A Script for Conversational Reflection Guidance'
[AY24]: Arakawa et al. (2024), 'Coaching copilot: blended form of an LLM-powered chatbot and a human coach to effectively support self-reflection for leadership growth'

2.3 Method

We present the *Interactive-Reflective Dialogue Alignment* (IRDA) system, which enables non-expert users to iteratively define a value concept and construct a corresponding reward model for agent training. Our approach is founded on the insight that preferences are noisy representations of values constructed and refined through reflection and iterative feedback [LS06]. To this end, IRDA employs a dual-loop framework that first elicits user feedback through a *preference construction loop* over a diversity-based pool of trajectories and then refines the model via an *uncertainty reduction loop* over a separate pool.[4]

[LS06]: Lichtenstein et al. (2006), 'The construction of preference: An overview'

4: The preference construction loop can be seen as reducing the user's (human's) uncertainty about how their values apply in the given context, while the uncertainty reduction loop can be thought of as reducing the reward model's uncertainty regarding the user's clarified preferences.

To begin, let $\mathcal{T}_D$ denote the diversity-based pool of trajectories, where each trajectory $\tau = (s_0, a_0, \ldots, s_T)$ represents the sequence of states and actions executed by the agent. We encode the trajectory in a vector of numerical features $\phi(\tau) \in \mathbb{R}^d$ and partition $\mathcal{T}_D$ into k clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ using k-means clustering:

$$\{\mathcal{C}_i\}_{i=1}^{k} = \arg\min_{\{\mathcal{C}_i\}} \sum_{i=1}^{k} \sum_{\tau \in \mathcal{C}_i} \|\phi(\tau) - \mu_i\|_2^2,$$

with cluster centroids

$$\mu_i = \frac{1}{|\mathcal{C}_i|} \sum_{\tau \in \mathcal{C}_i} \phi(\tau).$$

For each cluster, we select a representative trajectory:

$$\tau_i^{cent} = \arg\min_{\tau \in \mathcal{C}_i} \|\phi(\tau) - \mu_i\|_2.$$

The goal of this clustering is to improve learning efficiency. By selecting representative trajectories that cover a broad range of possible agent behaviors, the system aims to elicit a comprehensive understanding of the user's preferences from a minimal set of examples.

In the initial *preference construction loop*, the user first specifies a value concept, such as "respectfulness." The user is then presented, one at a time, with visualizations[5] of the k centroid trajectories, τ_i^{cent}. For each visualization, the user provides qualitative feedback e_i (e.g., "This is not respectful because..."). Concurrently, each centroid trajectory τ_i^{cent} is encoded into an ASCII representation $\alpha(\tau_i^{cent})$. An example of this encoding is shown in Figure 2.2, and the specific ASCII encodings used in the experiments are detailed in Section A.1 of the appendix. The system aggregates the feedback by pairing the ASCII representation of each centroid trajectory with the corresponding user feedback, resulting in the dataset:

$$\mathcal{D}_{fb} = (\alpha(\tau_i^{cent}), e_i)_{i=1}^{k}.$$

Once the feedback $\mathcal{D}_{fb}$ is collected, an LLM is queried using a specific prompt (see Section A.1) and the entire feedback dataset $\mathcal{D}_{fb}$. The LLM processes this input to generate two outputs: first, a hypothesis $\mathcal{H}$ about the features the user is using to make their decisions, and second, a set of alternative features $\mathcal{A}$ the user could consider in their decision making. The primary purpose of presenting $\mathcal{A}$ is to encourage user reflection by offering different perspectives on the decision-making process, drawing inspiration from principles of deliberative democracy [Mur21].[6] A secondary benefit is that $\mathcal{A}$ may identify relevant features that were not represented in the specific trajectories initially presented to the user, $\mathcal{T}_D$.

The user is then asked to respond to the generated hypotheses and alternatives, explaining why these features are or are not significant to their decision-making. This is intended to help the user reflect on their values, which can update their mental model $\mathcal{M}_u$. If $\mathcal{M}_u$ does change, the user returns to the beginning of the *preference construction loop*.

Once the user confirms that $\mathcal{M}_u$ is stable, the system transitions to the *uncertainty reduction loop*. In this stage, we consider a separate uncertainty-based pool of trajectories, $\mathcal{T}_U$. We iteratively refine our reward model on this pool. The reward model is based on an LLM that is prompted to assess whether a trajectory is aligned or not (the prompt is given in Section A.1). The LLM is given the entire conversation history $\mathcal{C}$ (which includes all user feedback, system prompts, and responses), the encoded trajectory $\alpha(\tau)$, and environment details such as symbol meanings and action spaces (EnvDesc), then outputs token probabilities for labels such as "respectful" and "disrespectful." Specifically, the alignment probability $p_\theta(1|\tau)$ and misalignment probability $p_\theta(1|\tau)$ are computed as

$$p_\theta(1|\tau), p_\theta(0|\tau) = f_{\text{LLM}}\big(\text{EnvDesc}, \mathcal{C}, \alpha(\tau)\big),$$

where $p_\theta(1|\tau)$ is the token probability for the "aligned" token (e.g., "respectful") and $p_\theta(0|\tau)$ is the token probability for the "misaligned" token (e.g., "disrespectful").

```
----- Step: 0 -----
The main agent (M) is in the
quadrant (orchard) that it owns.
 M   .   . | .   G   .
 A   .   . | G   .   .
 .   .   . | .  AA   B
----------------------
 .   .   . | .   .   .
 B   .   A | B   .   .
 A   .   . | A   A   .
======================
----- Step: 1 -----
Main agent moved from [0, 0] to
     [1, 0].
 .   M   . | .   G   .
 A   .   . | G   .   .
 .   .   . | .  AA   B
----------------------
 .   .   . | .   .   .
 B   .   A | B   .   .
 A   .   . | A   A   .
```

Figure 2.2: ASCII encoding of two timesteps of a trajectory of the multi-agent apple farming environment.

5: Either an image or video, depending on the setting.

[Mur21]: Muradova (2021), 'Seeing the other side? Perspective-taking and reflective political judgements in interpersonal deliberation'

6: Presenting alternative features aims to stimulate reflection by encouraging the user to engage in perspective-taking—actively considering others' viewpoints and experiences. For an analysis of perspective-taking as a mechanism fostering reflective judgments in deliberative settings, see Muradova [Mur21].

Algorithm 1 *Interactive-Reflective Dialogue Alignment*

Input: $\mathcal{T}_D$, $\mathcal{T}_U$, value v, threshold ϵ, EnvDesc
Preprocessing: For each $\tau \in \mathcal{T}_D$, extract $\phi(\tau)$; cluster via k-means; select representatives $\{\tau_i^{cent}\}$.
repeat *Preference Construction Loop*
  For each τ_i^{cent}, obtain/update label e_i; form

$$\mathcal{D}_{fb} = \{(\alpha(\tau_i^{cent}), e_i)\}.$$

  Query the LLM with $\mathcal{D}_{fb}$ to yield feature hypotheses $\mathcal{H}$ and alternatives $\mathcal{A}$.
  User responds to $(\mathcal{H}, \mathcal{A})$ and refines $\mathcal{M}_u$.
until User confirms $\mathcal{M}_u$ is stable
repeat *Uncertainty Reduction Loop*
  for each $\tau \in \mathcal{T}_U$ **do**
    $p_\theta(1|\tau), p_\theta(0|\tau) = f_{\text{LLM}}(\text{EnvDesc}, \mathcal{C}, \alpha(\tau))$
    $U(\tau) = 1 - \left|p_\theta(1|\tau) - p_\theta(0|\tau)\right|$
  end for
  $\tau^* = \arg\max_{\tau \in \mathcal{T}_U} U(\tau)$.
  Query the user for e^* on τ^*; update

$$\mathcal{D}_{fb} \leftarrow \mathcal{D}_{fb} \cup \{(\alpha(\tau^*), e^*)\}.$$

until $U(\tau^*) < \epsilon$
Final Reward Model: For any new τ, define

$$R(\tau) = \mathbb{I}\left[p_\theta(1|\tau) > p_\theta(0|\tau)\right],$$

Return: $R(\cdot)$.

The associated uncertainty is defined by

$$U(\tau) = 1 - \left|p_\theta(1|\tau) - p_\theta(0|\tau)\right|,$$

and the trajectory with maximum uncertainty is selected,

$$\tau^* = \arg\max_{\tau \in \mathcal{T}_U} U(\tau),$$

and the user is queried for an explanation e^* regarding τ^*. This new feedback $(\alpha(\tau^*), e^*)$ is appended to $\mathcal{D}_{fb}$, and the uncertainty reduction loop is repeated until $U(\tau^*) < \epsilon$ for a set threshold ϵ.

The final reward model is an LLM that, using the complete conversation history $\mathcal{C}$ and all accumulated feedback $\mathcal{D}_{fb}$, classifies new trajectories as aligned or misaligned with the user's value. It outputs a binary decision based on token probabilities:

$$R(\tau) = \mathbb{I}\left[p_\theta(1|\tau) > p_\theta(0|\tau)\right].$$

This unified process—beginning with diversity-based sampling, followed by preference construction, and culminating in uncertainty-driven refinement yields a final LLM-based reward model. This final LLM-based reward model leverages the entire conversation history and all explained examples. The full process is formalized in Algorithm 1.

2.4 Study Design

We evaluated our system in two studies: Study 1 investigates the utility of our system for learning about participants' definition of *respectful* agent behavior. Study 2 investigates the utility of our system for learning about participants' decision-making in moral dilemmas involving an agent (autonomous vehicle). Our studies employ a within-subject design, collecting data from each participant to train and test each method. Our studies aim to answer the following three questions:

RQ1: How do individuals' interpretations of value-aligned AI behavior differ?
RQ2: Does structured reflection enhance language-based reward modeling?
RQ3: When is individualized language-based reward modeling effective?

2.4.1 Environments

Multi-Agent Apple Farming Environment (Study 1). A 6×6 grid contains apples and garbage, with one "main" (blue) agent and three "background" (grey) agents. Each agent "owns" one of four 3×3 orchards. Two background agents remain stationary, and one moves freely. The main agent is rewarded by picking apples (none for collecting garbage). Participants assess whether the blue agent behaves "respectfully."

Moral Machine Environment (Study 2). Adapted from the Moral Machine experiment by Awad et al. [Awa+18], this environment presents users with a series of ethical dilemmas faced by an autonomous vehicle (AV). Each scenario involves a choice between two unavoidable outcomes: one resulting from the AV staying on its current course and the other from the AV swerving. Importantly, any decision results in the fatalities of one or more characters. The outcomes presented in each dilemma vary based on combinations of nine manipulated attributes:

[Awa+18]: Awad et al. (2018), 'The moral machine experiment'

- **Structural Attributes:** Intervention type (staying course vs. swerving), the relationship of potential casualties to the AV (passengers vs. pedestrians), and legality (whether pedestrians are crossing lawfully or unlawfully).
- **Character Attributes:** Gender (male vs. female), age (younger vs. older), social status (higher vs. lower), fitness (fit vs. large physique), the number of characters involved, and species (humans vs. pets, specifically dogs and cats).

These attributes were embodied by 20 distinct character types, including specific roles (e.g., doctor, executive, criminal, homeless person), age groups (e.g., elderly man, girl, baby stroller), gender, fitness levels (e.g., large man, female athlete), and animals (dog, cat). Participants were tasked with deciding which of the two outcomes the AV should choose in each dilemma. A full description of the scenario generation and attributes is available at `https://osf.io/wt6mc`.

2.4.2 Participants

In Study 1, we recruited 21 participants from the University of Waterloo (18 to 39 age range, M=23.86, 7 self-identified as male and 14 as female). When asked to rate their level of familiarity with reinforcement learning on a 5-point Likert ranging from "very unfamiliar" (1) to "very familiar" (5), the mean level of familiarity was 2.48, with the mode and median being 2. The Likert-scale data is visualized in figure 2.3 and highlights that more than half of participants were "unfamiliar" or "very unfamiliar" with reinforcement learning.

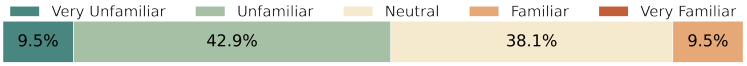


Figure 2.3: Participant familiarity with reinforcement learning in Study 1 of Chapter 2.

In Study 2, we recruited 9 participants from the University of Waterloo (18 to 33 age range, M=25.66, 6 self-identified as male and 3 as female). When asked to rate their level of familiarity with reinforcement learning on a 5-point Likert ranging from "very unfamiliar" (1) to "very familiar" (5), the mean level of familiarity was 3.55, with the mode and median being 3. The Likert-scale data is visualized in figure 2.4.



Figure 2.4: Participant familiarity with reinforcement learning in Study 2 of Chapter 2.

2.4.3 Procedure

Participants used our system to specify how they would like the agent to act. Participants then labeled 50 scenarios and were interviewed.

Introduction (~5min) - After participants completed the consent form and demographic questions, we thoroughly explained the environment mechanics so that differences reflected their opinions rather than assumptions about the setup.

Dialogue - The participant began by conversing with the system about the agent's behavior following the process described in Algorithm 1. To control the amount of time users spent, we limited the user to one preference construction loop and one uncertainty reduction loop.

Labeling - Following the participants' dialogue interaction with the system, participants labeled 20 scenarios. Each participant labeled the same scenarios, which allowed us to assess how much the participants agreed on the labels.

Semi-structured Interview (~10min) - After completing the labeling task, participants were asked about their ability to communicate their decision-making to the system, including their capacity to articulate label choices, any difficulties in decision-making, and potential changes in their labeling behavior over time.

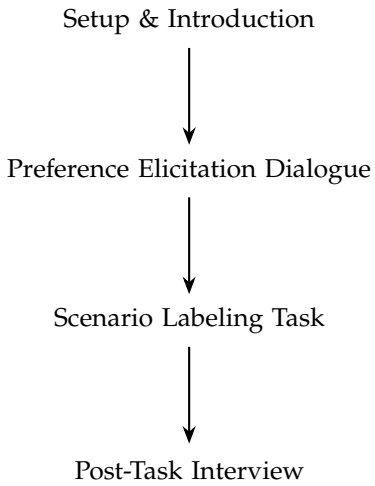


Figure 2.5: Study procedure for both studies in Chapter 2.

2.4.4 Baseline Comparisons

To evaluate the effectiveness and efficiency of IRDA, we compared its performance against several baseline methods. These baselines were chosen to isolate the specific contributions of IRDA's unique component, namely the reflective dialogue, and to benchmark its performance against standard machine learning techniques under varying data conditions.

Language-Based Baseline (L^B)

Kwon et al. [Kwo+23] proposed a reward modeling pipeline for text-based environments where the user selects multiple examples from a palette of examples of the agent behaving as they would desire, accompanied by explanations. We modify their pipeline in the following way: Instead of asking the user to select examples from a handcrafted palette, we choose the examples the user sees with the diversity- and uncertainty-based sampling procedures described and in Section 2.3. This system differs from IRDA in that it does not engage the user in reflective dialogue.

[Kwo+23]: Kwon et al. (2023), 'Reward Design with Language Models'

The primary purpose of the L^B baseline is to serve as a direct comparison to IRDA, differing only in the absence of the structured reflective dialogue component (specifically, the generation and discussion of feature hypotheses $\mathcal{H}$ and alternatives $\mathcal{A}$). By comparing IRDA against L^B, we can directly assess the impact of the reflective interaction on the quality of the learned reward model, thereby addressing **RQ2** (Does structured reflection enhance language-based reward modeling?). It allows us to isolate the contribution of the dialogue mechanism beyond simply using an LLM to process trajectory-explanation pairs.

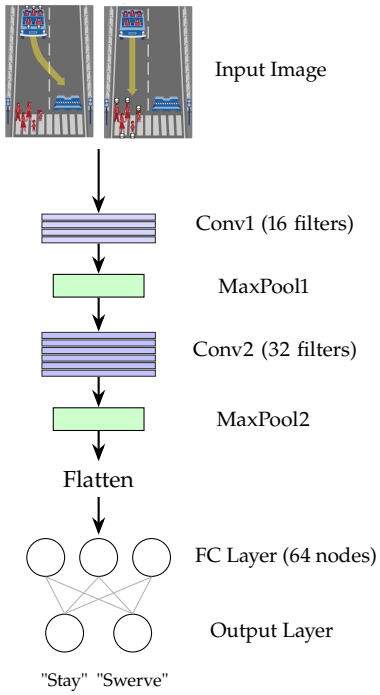


Figure 2.6: CNN architecture used in Study 2 of Chapter 2: Feature Extraction (Conv/Pool layers) feeds into Classification (FC layers). Output neurons for 'Stay' and 'Swerve.'

Supervised Learning Baselines

We also compared IRDA against traditional supervised learning approaches, which represent standard methods for learning predictive models from labeled data. These comparisons help benchmark IRDA's performance and, crucially, its sample efficiency. We implemented two types of supervised models: Individual models per participant and collective models trained on aggregated data. In Study 1, we implemented multi-layer perceptron models (one hidden layer, 32 neurons) using a tensor encoding of the trajectory: individual models ($\text{MLP}_i^{\text{ind}}$) for each participant i and a collective model (MLP^{col}) trained on all participant data. Study 2 expanded this comparison to include both MLP models with 26-dimensional Moral Machine scenario vectors and convolutional neural networks ($\text{CNN}_i^{\text{ind}}$ and CNN^{col}) using scenario image inputs. The CNNs used two convolutional layers (16 and 32 filters) with max pooling, followed by fully connected layers reducing to 64 dimensions. Each supervised learning model was incrementally trained with up to 30 examples *per participant* using the Adam optimizer (learning rate 0.001).

Table 2.1: Mapping of Analysis Methods to Research Questions

Analysis Method	RQ1	RQ2	RQ3
Inter-Annotator Agreement	✓		✓
Evaluation of Language-Based Reward Model Performance		✓	
Comparison to Supervised Learning	✓		✓
Qualitative Analysis of Participant Decision Making	✓		
Analysis of Feature Similarity Between Participants	✓		✓
Interview Thematic Analysis		✓	✓

2.4.5 Analysis

To answer our research questions, we employ a mixed-methods approach, combining quantitative analyses of model performance and inter-annotator agreement with qualitative analyses of participant decision-making processes and experiences.

Inter-Annotator Agreement

We assess the inter-annotator agreement between participants on the test set of scenarios they labelled in each study. Since each participant labelled the same test scenarios, we can use Fleiss' kappa value to quantify the inter-annotator agreement between the participants [LK77]. Generally, kappa statistics below 0 indicate "poor" agreement and kappa statistics above 0.8 indicate "nearly perfect" agreement.

[LK77]: Landis et al. (1977), 'The measurement of observer agreement for categorical data'

Evaluation of Language-Based Reward Model Performance

We evaluated our system against a baseline without dialogic reflection using a performance metric P, where P_i^{IRDA} and P_i^B represent participant i's metrics for our system and baseline, respectively. Study 1 used balanced accuracy due to class imbalance, while Study 2 used accuracy. For each participant, both systems generated rewards for 20 non-training scenarios, yielding 20 pairs of P values. We conducted three statistical tests on the P values:

1. We bootstrapped 95% confidence intervals for the mean by resampling 10,000 times with replacement.
2. For each participant, we calculated the difference $\Delta P_i = P_i^{\text{IRDA}} - P_i^B$ and bootstrapped these differences in the same way.
3. P values were compared using the Wilcoxon signed-rank test, chosen for its robustness to non-normal distributions and reduced false positives [BS99].

[BS99]: Bridge et al. (1999), 'Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research'

Comparison to Supervised Learning

We compared our language-based systems to traditional supervised learning approaches. Both the individual models ($\text{MLP}_i^{\text{ind}}$ and $\text{CNN}_i^{\text{ind}}$) and the collective models (MLP^{col} and CNN^{col}) were trained incrementally, gradually increasing the number of samples used per participant. This methodology allowed us to analyze how model performance evolved with increasing data availability. For each increment, we calculated P_i^{ind}

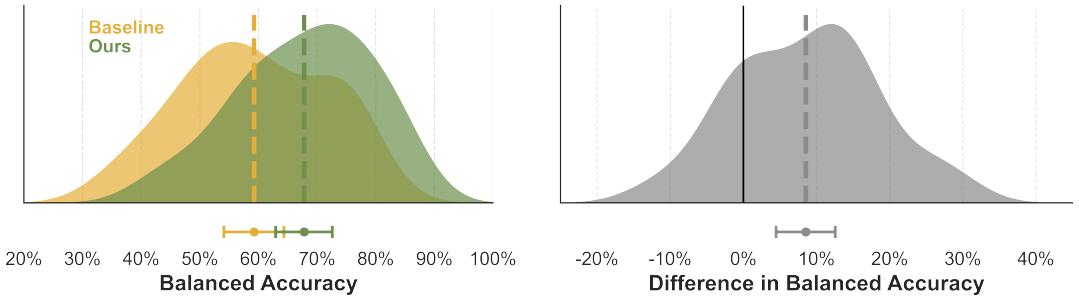


Figure 2.7: (Left) Distributions of balanced accuracies for language-based reward models: our pipeline (IRDA) vs. baseline (L^B) in Study 1 of Chapter 2. (Right) Distribution of per-participant differences in balanced accuracy ($P_i^{\text{IRDA}} - P_i^B$) between IRDA and baseline models in Study 1 of Chapter 2.

and P_i^{col} for each participant i. To ensure robustness, we bootstrapped these values with replacement using 10,000 resamples.

Qualitative Analysis of Participant Decision Making

We conducted a detailed analysis of the message exchanges between participants and the system to gain insight into participants' decision-making processes. We employed an inductive coding approach, systematically reviewing the messages to identify key features and criteria that participants used in their decision-making. Our coding process involved multiple passes through the data, with iterative refinement of the codebook to ensure it captured the full range of decision-making strategies observed.

Analysis of Feature Similarity Between Participants

To quantify how similar participants were in their use of decision-making features, we employed the Jaccard similarity coefficient [Jac12]. This measure calculates the overlap between two sets of items, which, in our case, are features that the two participants used to make decisions. We computed the Jaccard similarity coefficient for every pair of participants, using the set of decision-making features each participant employed (as identified in our qualitative analysis). To estimate the overall similarity across our participant pool, we then calculated the mean of these pairwise Jaccard coefficients. We used bootstrapping with 10,000 resamples to determine the 95% confidence intervals.

[Jac12]: Jaccard (1912), 'The distribution of the flora in the alpine zone'

Thematic Analysis of Interview Data

We conducted semi-structured interviews with participants to understand their experiences. The interview transcripts were analyzed using a thematic analysis approach guided by the principles outlined by Braun and Clarke [BC06]. We followed a six-phase process: familiarization with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report.

[BC06]: Braun et al. (2006), 'Using thematic analysis in psychology'

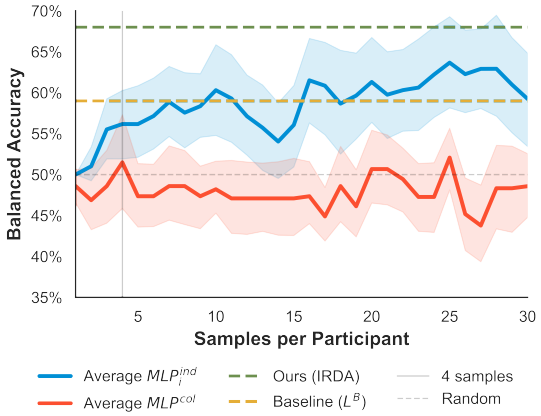


Figure 2.8: Balanced accuracy of models vs. samples per participant in **Study 1** of Chapter 2. Blue line shows average individual MLP (MLP^{ind}); red shows collective MLP (MLP^{col}). Our IRDA system (green dashed) and baseline (L^B, yellow dashed) used 4 samples per participant (vertical gray line). Collective model trained on 21x samples shown (21 participants). Shaded areas: 95% confidence intervals. Gray dashed line: random performance.

2.5 Results: Study 1 - Multi-Agent Apple Farming

On average, participants took 15 minutes 57 seconds (SD = 6 min. 43 sec., range: 6 min. 59 sec. - 30 min. 55 sec.) to complete the dialogue with the system and 13 minutes 37 seconds (SD = 3 min. 2 sec., range: 6 min. 55 sec. - 18 min. 26 sec.) to complete the labelling of 50 trajectories. Of 21 participants, 7 (33.$\bar{3}$3%) entered the *preference construction loop*[7].

7: As noted above, participants were limited to one iteration of the preference construction loop due to time constraints.

S1 - Inter-Annotator Agreement

We observed a Fleiss' kappa value between all participants' labels on the 50 labelled trajectories of $\kappa = 0.336$, indicating "fair" agreement among participants [LK77]. The Fleiss' kappa statistic of 0.336 we observed lends credence to the idea that human values and preferences are subjective and personal.

[LK77]: Landis et al. (1977), 'The measurement of observer agreement for categorical data'

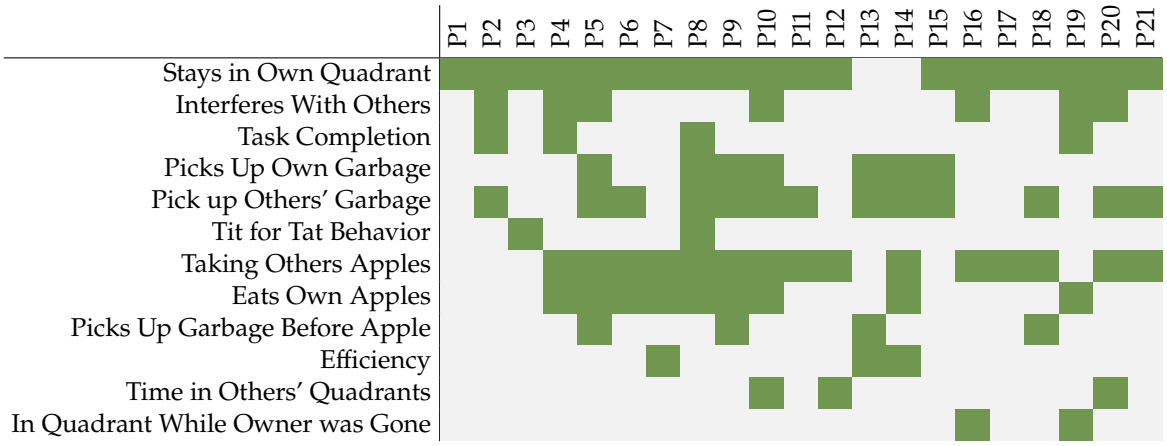
S1 – Evaluation of Language-Based Reward Model Performance.

On average, the reward models produced by our pipeline (IRDA) received significantly higher balanced accuracy scores (measured in percentages) than the baseline system (L^B) by 9% (95% CI: [5%, 13%], M = 68% vs. M = 59%, p=.002). This indicates that structured reflection is beneficial. The distributions of the balanced accuracies for each pipeline are visualized in the left frame of Figure 2.7, and the distribution of the per participant difference in balanced accuracy is shown in the right frame.

S1 – Comparison to Supervised Learning

With all 30 training samples, the average balanced accuracy of the individual models (MLP_i^{ind}) was 59% (95% CI: [53%, 65%]) while the collective model (MLP^{col}) achieved 48% (95% CI: [46%, 50%]). This indicates that participant value definitions varied significantly. Figure 2.8

Table 2.2: Behavioural features participants used to decide whether the agent was acting respectfully in Study 1 of Chapter 2. Cells coloured in light grey ( ) indicate that the participant did not consider the feature. Cells colored in green ( ) indicate the participant considered the feature in their decision making.



illustrates the relationship between model performance and the number of samples provided per participant.

S1 – Qualitative Analysis of Participant Decision Making

Although our system can align AI agents with various values, we focused on respect to examine how individuals interpret even a single value differently. Analysis of participant conversations revealed 12 distinct behavioral features used to evaluate respectful agent behavior.[8] Usage varied significantly - P1 relied solely on whether agents stayed in their quadrant, while P5, P7, and P10 each employed seven features. Only one participant pair shared identical feature sets, and most participants combined them using hierarchical and conditional rules. While the agent staying in its quadrant was the most common feature, features varied in temporal scope, from static properties (current quadrant location) to multi-step sequences (collecting garbage before apples). A full list and description of the features we identified can be found in Section A.2 in the appendix.

8: We analyzed features only from the five trajectories (out of six) that all participants viewed.

To convey the diversity of decision making features participants used, we visualized the feature use in Table 2.2.

S1 – Analysis of Feature Similarity Between Participants.

We observed an average Jaccard similarity coefficient between all pairs of participants' feature usage of $J = 0.357$, with a 95% confidence interval of $(0.333, 0.3813)$.

S1 – Thematic Analysis of Interview Data

Our thematic analysis revealed two main themes: participants' evolving definitions of respect and the system's impact on this evolution.

Evolving Definitions of Respect. Participants' understanding of respect developed significantly through system interaction. Initial definitions

focused on simple concepts like spatial boundaries and task-specific behaviors (P4, P6, P7, P10, P19). Through engagement with examples and system feedback, these views became more nuanced. Participants who encountered examples challenging their initial perspectives often expanded their conceptualization of respect (P3, P8, P10, P13, P18, P19, P20). This evolution aligns with consumer research showing that reflection and realistic decision-making improve preference reporting [HDD14].

[HDD14]: Hauser et al. (2014), 'Self-reflection and articulated consumer preferences'

Specific Impact of System Hypothesis. The system's hypotheses and alternative features actively shaped participants' respect definitions. The presentation of alternatives prompted reevaluation and revision of initial concepts (P3, P8, P13, P18, P20). While some participants maintained their original views (e.g., P19), the system's suggestions helped others refine their understanding (P13, P18) or consider new perspectives (P3, P20), demonstrating the value of reflective dialogue. This finding corroborates similar findings in group-based deliberation [Mur21].[9]

[Mur21]: Muradova (2021), 'Seeing the other side? Perspective-taking and reflective political judgements in interpersonal deliberation'

9: As discussed in Chapter 1, we hypothesized that LLMs could aid individuals in forming preferences, similar to the process observed in deliberative settings like citizens' assemblies. The findings of this study are consistent with this hypothesis. Exploring the extent and conditions under which LLMs can serve this function presents a valuable avenue for future research.

2.6 Results: Study 2 - The Moral Machine

On average, participants took 18 minutes 28 seconds (SD = 7 min. 15 sec., range: 12 min. 0 sec. - 34 min. 34 sec.) to complete the dialogue with the system and 11 minutes 51 seconds (SD = 4 min. 15 sec., range: 4 min. 46 sec. - 17 min. 04 sec.) to complete the labelling of 50 trajectories. Of 9 participants, 1 (11.$\bar{1}$1%) entered the *preference construction loop* for one iteration.

S2 - Inter-Annotator Agreement

We observed a Fleiss' kappa value between all participants' labels on the 50 labeled trajectories of $\kappa = 0.460$, indicating "moderate" (higher than "fair") agreement among participants [LK77].

[LK77]: Landis et al. (1977), 'The measurement of observer agreement for categorical data'

S2 – Evaluation of Language-Based Reward Model Performance.

On average, the reward models produced by our pipeline (IRDA) received significantly higher accuracy scores (measured in percentages) than the baseline system (L^B) by 12% (95% CI: [4%, 27%], M = 65% vs. M = 53%, p=.05). This adds more evidence in favor of the effectiveness of structured reflection. The distributions of the balanced accuracies for each pipeline are visualized in the left frame of Figure 2.9. The distribution of the per-participant difference in balanced accuracy is shown in the right frame.

S2 – Comparison to Supervised Learning

With all 30 training samples, the average accuracy of the individual MLP models ($\text{MLP}_i^{\text{ind}}$ was 79% (95% CI: [74%, 84%]) while the collective model (MLP^{col}) achieved 77% (95% CI: [75%, 78%]). The left frame Figure 2.10 illustrates the relationship between model performance and the number of samples provided per participant for the MLP models. For the CNN models, with all 30 training samples, the average accuracy

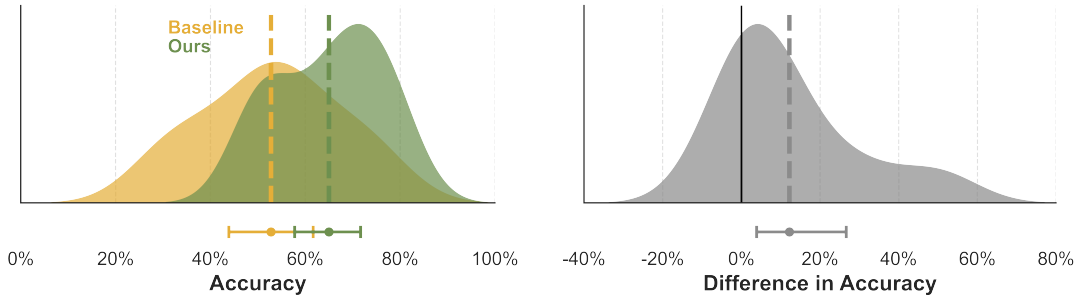


Figure 2.9: (Left) Distributions of accuracies for language-based reward models: our pipeline (IRDA) vs. baseline in Study 2. (Right) Distribution of per-participant differences in accuracy ($P_i^{\text{IRDA}} - P_i^{B}$) between IRDA and baseline models in Study 2.

of the individual models ($\text{CNN}_i^{\text{ind}}$) was 67% (95% CI: [61%, 73%]) while the collective model (CNN^{col} achieved 77% (95% CI: [70%, 83%]). The right frame of Figure 2.10 illustrates the relationship between model performance and the number of samples provided per participant for the CNN models. These results point to two things: first, when participant agreement is high, collective methods may outperform individualized methods. Second, when agreement is high and the learning problem becomes more difficult (e.g., the CNN models with image input), pooling of samples is beneficial.

S2 – Qualitative Analysis of Participant Decision Making

By analyzing the participants' conversations with our system, we identified nine features they used in their decision-making in Study 2. The most common features were minimizing casualties (8/9 participants) and traffic rule compliance (8/9 participants). Most participants combined features conditionally and hierarchically. The features each participant used are visualized in Table 2.3. Descriptions of each feature can be found in Section A.2 in the appendix.

	P1	P2	P3	P4	P5	P6	P7	P8	P9
Human vs. Animal Priority	■	■	■			■	■		■
Traffic Rule Compliance	■	■	■	■	■	■		■	■
Minimizing Casualties	■	■	■	■	■	■	■		■
Age Prioritization	■		■		■	■	■		■
Survival Probability	■				■	■			
Maintaining Original Path		■	■	■				■	■
Societal Value/Contribution					■	■	■	■	■
Passenger Safety Priority	■			■					
Gender							■		

Table 2.3: Behavioural features participants considered (Study 2). Cells coloured in light grey ( ) indicate that the participant did not consider the feature. Cells colored in green (■) indicate the participant considered the feature.

S2 – Analysis of Feature Similarity Between Participants.

We observed an average Jaccard similarity coefficient between all pairs of participants' feature usage of $J = 0.464$, with a 95% confidence interval of $(0.403, 0.526)$.

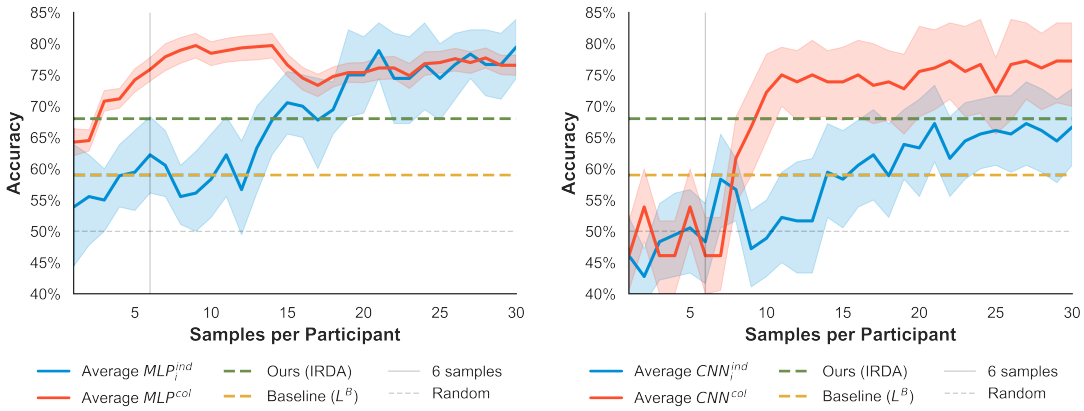


Figure 2.10: Comparison of model accuracies as a function of samples per participant in Study 2. (Left) Performance of MLP-based models: average individual MLP (MLP^{ind}, blue), collective MLP (MLP^{col}, red), our IRDA approach (green dashed), and baseline L^B (yellow dashed). (Right) Performance of CNN-based models: average individual CNN (CNN^{ind}, blue), collective CNN (CNN^{col}, red), IRDA, and baseline. Both panels show confidence intervals (shaded areas), the 6-sample training point for IRDA and baseline (vertical gray line), and the random performance level (gray dashed).

S2 – Thematic Analysis of Interview Data

Through our thematic analysis, we found two main themes: (1) participants' definitions of ethical decision making evolved throughout the activity, and (2) participants' decisions were primarily based on explicit reasoning but sometimes relied on intuition.

Decision-making Evolution. The interviews showed variations in how participants' decision-making approaches developed throughout their interaction with the system and the Moral Machine scenarios. For some, the process led to an evolution in their ethical criteria. These participants described refining or expanding their decision-making rules as they encountered a wider range of dilemmas. For instance, P3 noted that exposure to scenarios where their initial criteria resulted in a tie forced them to "realize[...] the need to consider new factors." This adaptation, triggered by the complexity and diversity of the ethical trade-offs presented, mirrors the findings on preference construction and evolution observed in Study 1, where engagement with challenging examples prompted deeper reflection.

Conversely, other participants reported maintaining a consistent ethical framework throughout the study. P2, for example, stated their "rules remained consistent," suggesting they approached the task with a predefined set of principles that were applied across different scenarios. For these participants, the interaction may have served less as a process of preference construction and more as an opportunity to articulate and apply their existing, stable values to the specific dilemmas presented by the Moral Machine environment. This highlights that while the reflective process can facilitate preference evolution for some, it can also help others express pre-existing convictions.

Intuition vs. Explicit Reasoning. While most participants could articulate clear reasoning, some relied on intuition for complex scenarios. P7 reported using "first instinct" or "vibes" for several challenging cases.

2.7 Discussion

The *Interactive-Reflective Dialogue Alignment* (IRDA) system, developed as part of this thesis's exploration of the interface between Artificial Intelligence and Social Choice theory, yields insights relevant both for aligning AI with diverse human values (the **SC** → **AI** direction) and for leveraging AI to improve preference elicitation and collective decision-making processes (the **AI** → **SC** direction). Our evaluation suggests significant value diversity exists in AI alignment contexts and that LLM-guided reflection, as implemented in IRDA, is an effective, sample-efficient method for learning individualized reward models that capture this pluralism.

Value Diversity (RQ1). Our studies revealed context-dependent value pluralism. In Study 1 (respectful behavior), participants exhibited substantial disagreement ($\kappa = 0.336$, $J = 0.357$), and collective models performed poorly compared to individualized ones. This heterogeneity underscores the potential inadequacy of standard alignment methods that implicitly aggregate preferences, motivating the need for frameworks grounded in social choice principles (**SC** → **AI**). Such frameworks often require explicit, individualized preference representations—like those IRDA generates—as a necessary foundation for applying transparent and principled aggregation rules. In contrast, Study 2 (Moral Machine) showed higher agreement ($\kappa = 0.460$, $J = 0.464$), where collective models were more competitive. This suggests that the degree of value pluralism is context-specific, influencing the choice of appropriate alignment or aggregation techniques.

The Effectiveness of Reflection (RQ2). IRDA significantly outperformed the non-reflective baseline (L^B) in both studies (9-12% accuracy improvement), demonstrating the benefit of structured reflection for preference elicitation. This supports the thesis argument (Chapter 1) that LLMs can enhance collective decision-making processes (**AI** → **SC**) by moving beyond the limitations of traditional methods like polls or ballots, which often capture "top-of-the-head" responses [FL05]. By facilitating reflection, IRDA helps users engage in preference *construction* [Fis91], articulating more considered judgments and translating latent values into operational reward signals useful for alignment. The improvement, even among participants with stable views, suggests reflection enhances articulation and communication clarity.

[FL05]: Fishkin et al. (2005), 'Experimenting with a democratic ideal: Deliberative polling and public opinion'

[Fis91]: Fischhoff (1991), 'Value elicitation: Is there anything in there?'

Contextual Efficacy of Language-Based RMs (RQ3). The relative advantage of individualized, language-based reward modeling depends on context. IRDA's strength lies in its sample efficiency and ability to capture heterogeneous preferences (Study 1), leveraging LLM few-shot learning and reflection. However, with more homogeneous preferences and sufficient data, traditional collective models may match or exceed the performance of individualized models trained on sparse data (Study 2). This contextual variation informs the practical application of methods for both pluralistic alignment (**SC** → **AI**) and AI-assisted decision-making (**AI** → **SC**). It suggests when the richer, individualized preference information captured by IRDA is most critical, versus situations where simpler collective modeling might be adequate or even preferable due to robustness with larger datasets.

Contributions and Implications. In summary, this chapter addressed challenges central to the thesis: handling value pluralism and improving preference elicitation through AI. We introduced and evaluated IRDA, a novel pipeline combining LLM-guided reflection, active learning, and in-context reward modeling. IRDA contributes directly to the **AI** $\rightarrow$ **SC** direction by offering an interactive, conversational technique for eliciting richer, more reflective verbal preferences than traditional methods allow. Simultaneously, by generating nuanced, individualized reward models, it provides the essential building blocks needed for applying explicit social choice aggregation mechanisms within the **SC** $\rightarrow$ **AI** direction, enabling more principled pluralistic alignment. Our empirical results characterized the nature of value divergence (RQ1), demonstrated the utility of reflection (RQ2), and mapped the conditions under which individualized modeling excels (RQ3). Together, these contributions advance the development of AI systems and AI-assisted processes designed to be sensitive to human diversity, empowering individuals to better articulate and embed their values in socio-technical systems.

Consensus Statement Generation via Token-Level Tree Search

3

Current frameworks for large language model generation lack the inherent structure needed to provide provable fairness guarantees when aggregating diverse free-form opinions into consensus statements. We model the task as a multi-objective, token-level Markov Decision Process (MDP), where each objective corresponds to an agent's preference. Token-level rewards for each agent are derived from their policy (e.g., a personalized language model). This approach utilizes the finding that such policies implicitly define optimal Q-functions, providing a principled way to quantify rewards at each generation step without a value function [Raf+24]. This MDP formulation creates a formal structure amenable to analysis using principles from social choice theory. We propose two approaches grounded in social choice theory. First, we propose a stochastic generation policy guaranteed to be in the ex-ante core, extending core stability concepts from voting theory to text generation. This policy is derived from an underlying distribution over complete statements that maximizes proportional fairness (Nash Welfare). Second, for generating a single statement, we target the maximization of egalitarian welfare using search algorithms within the MDP framework. Empirically, experiments using language models to instantiate agent policies show that search guided by the egalitarian objective generates consensus statements with improved worst-case agent alignment compared to baseline methods, including the Habermas Machine [Tes+24].

3.1 Motivation 22
3.2 Related Work 23
3.3 Problem Setup & Preliminaries 25
3.4 Stochastic Policies and Ex-Ante Fairness 28
3.4.1 Computational Tractability via Token Chunking . 29
3.4.2 Deriving the Ex-Ante Fair Stochastic Policy 29
3.4.3 Properties of the Induced Policy 30
3.4.4 A Note on Computation and Approximation . . . 32
3.5 Generating a Single Consensus Statement . . 32
3.5.1 The Rawlsian Path 33
3.6 Experiments 35
3.6.1 Evaluating Prompt-Based Credit Assignment 35
3.6.2 Consensus Generation . . 36
3.7 Discussion 38

3.1 Motivation

Collective deliberation, involving the exchange of diverse views to reach common ground, is fundamental to democratic processes. While structured forums like citizens' assemblies facilitate this, they face challenges of scale, cost, and ensuring equal participation [Tes+24]. Artificial intelligence, particularly Large Language Models (LLMs), can support deliberation by synthesizing complex viewpoints expressed in natural language. However, a gap exists: the lack of **principled mechanisms** for fairly aggregating these diverse perspectives during the generation of a consensus statement. Without transparent and theoretically grounded procedures, LLM-generated outputs risk obscuring disagreements or unfairly representing contributing voices.

[Tes+24]: Tessler et al. (2024), 'AI can help humans find common ground in democratic deliberation'

Social choice theory offers formal fairness criteria. The emerging field of *generative social choice* (GSC) aims to apply these principles to open-ended generation tasks [Fis+24; Tes+24]. Existing GSC approaches have targeted axioms like Balanced Justified Representation (BJR), which seeks fair representation for groups. However, methods to achieve these goals frequently rely on applying criteria post-hoc or using direct prompting strategies—for instance, instructing an LLM to generate text that maximizes the minimum utility [Fis+24]. While valuable, such approaches

[Fis+24]: Fish et al. (2024), 'Generative Social Choice'
[Tes+24]: Tessler et al. (2024), 'AI can help humans find common ground in democratic deliberation'

[Fis+24]: Fish et al. (2024), 'Generative Social Choice'

lack a step-by-step, verifiable mechanism operating *within* the generation process itself, making it difficult to assess the fairness of the final output.

This work introduces a framework that integrates fairness criteria into the token-level generation process, providing such a principled mechanism. We represent each agent i's viewpoint using a policy π_i, which assigns likelihoods $\pi_i(s, a)$ reflecting their desired content given the current prefix s. Text generation is then modeled as a deterministic, token-level Markov Decision Process (MDP). Within this MDP, agent rewards $r_i(s, a)$ are derived from their respective policy likelihoods (e.g., $r_i^{\log}(s, a) = \beta \log \pi_i(s, a)$), consistent with preference-based reward formulations [Raf+24].

[Raf+24]: Rafailov et al. (2024), 'From r to Q^*: Your Language Model is Secretly a Q-Function'

This MDP structure allows us to develop mechanisms for different fairness paradigms. One approach defines a **stochastic generation policy** Π^*. This policy is derived by optimizing a distribution over complete statements to achieve *proportional fairness*, specifically by maximizing the Nash Welfare product. We prove that the resulting policy Π^* resides in the *ex-ante core*, ensuring stability against coalitional deviations and providing a method for fair randomization grounded in the generation process.

Alternatively, for applications requiring a single consensus statement, our framework targets the maximization of *egalitarian welfare* (EW), defined as $\min_i U_i^{\log}(X)$, thereby seeking the best outcome for the least satisfied agent. This objective aligns with the representation fairness goals seen in GSC approaches like BJR. However, distinct from methods relying on high-level prompts, we propose **constructive algorithms**—specifically, finite lookahead search and beam search—that optimize the EW objective directly through structured planning and search over the token-level MDP. This provides a transparent and analyzable mechanism for generating statements that fulfill egalitarian aims.

This paper presents the formal MDP setup, derives theoretical fairness properties for the resulting generation policies, introduces tractable search algorithms, and provides empirical evidence supporting the validity of the reward formulation and the effectiveness of the proposed search methods compared to baselines.

3.2 Related Work

Our research on generating fair consensus statements draws upon and extends ideas from generative social choice, guided decoding, and mechanism design for large language models (LLMs).

Generative Social Choice. This field seeks to apply computational social choice principles to open-ended generation tasks, leveraging LLMs. Works like Fish et al. [Fis+24] and Tessler et al. [Tes+24] focus on aggregating diverse preferences into representative textual outputs (slates or consensus statements). These approaches employ LLM-based preference modeling, utilizing different strategies to ensure fairness or representation. Some methods apply specific aggregation rules, such as Schulze

[Fis+24]: Fish et al. (2024), 'Generative Social Choice'

[Tes+24]: Tessler et al. (2024), 'AI can help humans find common ground in democratic deliberation'

voting, to rank or select among candidate texts after they have been generated (post-hoc). Others aim to produce outputs that directly satisfy representation axioms, like JR/BJR, attempting to achieve this by guiding the generation process with high-level instructions within prompts [Fis+24]. Revel and Penigaud [RP25] provide conceptual framing for such AI-facilitated collective judgment systems.

[Fis+24]: Fish et al. (2024), 'Generative Social Choice'
[RP25]: Revel et al. (2025), 'AI-Facilitated Collective Judgements'

While valuable, these methods often lack a fine-grained, verifiable mechanism that integrates fairness into the step-by-step construction of the text. Our work addresses this gap by modeling statement creation as a sequence of choices within a multi-objective MDP, where selecting each token is treated as a public decision [CFS17]. This token-level optimization offers a more formally grounded and transparent mechanism compared to post-hoc evaluations or high-level prompting, ensuring fairness properties are respected throughout the construction of the consensus statement.

[CFS17]: Conitzer et al. (2017), 'Fair public decision making'

Guided Decoding. These techniques aim to control LLM generation at inference time towards desired properties. Methods often involve search algorithms guided by auxiliary models or objectives. For instance, PPO-MCTS (Liu et al. [Liu+24]) uses the PPO value network to score partial sequences within a Monte-Carlo Tree Search, while MOD (Shi et al. [Shi+24]) and COLLAB (Chakraborty et al. [Cha+25]) focus on combining or switching between multiple pre-existing models/policies at decoding time based on explicit weights or implicit Q-functions. Like these methods, we employ search algorithms (bounded lookahead, beam search) to navigate the generation space and optimize objectives during decoding. Our use of agent policies to provide token-level scores (likelihoods as rewards) also aligns with the core idea of using quantitative signals to guide generation. However, we frame the problem as planning within a multi-objective MDP derived directly from the input agent policies, optimizing social choice objectives (PF, Egalitarian Welfare) within this framework, rather than relying on a single pre-trained value model or heuristics for combining separate, pre-aligned models. We also prove that one of our methods is ex-ante fair (operationalized with proportional fairness and the core) and another one of our methods provides an approximation of ex-post optimal egalitarian welfare.

[Liu+24]: Liu et al. (2024), 'Don't throw away your value model! Generating more preferable text with Value-Guided Monte-Carlo Tree Search decoding'
[Shi+24]: Shi et al. (2024), 'Decoding-time language model alignment with multiple objectives'
[Cha+25]: Chakraborty et al. (2025), 'Collab: Controlled Decoding using Mixture of Agents for LLM Alignment'

Mechanism Design for LLMs. This nascent area explores mechanisms for settings where multiple agents interact via LLMs. Dütting et al. [Due+24] design token-level auctions where bids influence generated distributions, analyzing incentive compatibility. Dubey et al. [Dub+24] design auctions for incorporating ads into LLM summaries, using "prominence" as an intermediate allocation variable. Common ground exists in the high-level goal of aggregating inputs from multiple agents (represented algorithmically/via LLMs) to produce a collective textual output. However, our work differs significantly in methodology and objective. We do not employ economic mechanisms like auctions, bids, or payments. Instead, we formulate the aggregation problem as a multi-objective optimization within an MDP, aiming to achieve a fair consensus based on social choice criteria, rather than allocating influence or generating content based on bids.

[Due+24]: Duetting et al. (2024), 'Mechanism design for large language models'
[Dub+24]: Dubey et al. (2024), 'Auctions with llm summaries'

3.3 Problem Setup & Preliminaries

Consider a setting with a finite set of agents $N = \{1, 2, \ldots, n\}$, each representing distinct viewpoints or interests regarding a specific **Issue**. The goal is to generate a consensus text statement reflecting these perspectives fairly. The inputs to the process include descriptions of the Issue, the **agent opinions** (e.g., free-form text expressing their views), and the derived **agent policies**. Additionally, a **reference consensus policy** (e.g., a base language model) is used to propose tokens in the consensus statement.

Agent Policies. Each agent $i \in N$ is represented by a *policy* π_i, which assigns a likelihood $\pi_i(s, a) \in [0, 1]$ to each action a given the state s. Intuitively, $\pi_i(s, a)$ reflects how closely an action aligns with agent i's preference at state s. This policy could be an LLM fine-tuned or prompted for agent i's viewpoint. We assume $\sum_{a \in A} \pi_i(s, a) = 1$ for all i, s.

Token-Level MDP. We model text generation as a deterministic, discrete-time Markov Decision Process defined by the tuple $(S, A, T, \mathbf{R})$. Here, S is the state space of partial text sequences (prefixes), including initial s_0 and terminal states. A is the action space consisting of the token vocabulary plus a special end-of-sequence token $\langle \text{eos} \rangle$. T is the deterministic transition function where $T(s, a) = s \| a$ appends the chosen token; selecting $a = \langle \text{eos} \rangle$ leads to a terminal state representing a completed statement X. Finally, $\mathbf{R}$ represents the agent-specific reward functions. We define two types of rewards based on agent policies, serving different analytical purposes:

1. **Log-Likelihood Reward:** $r_i^{\log}(s, a) = \beta \log \pi_i(s, a)$. This formulation aligns with implicit rewards in preference learning [Raf+24], where $\beta > 0$ is a scaling factor. This reward can be negative and is suitable for additive utility accumulation along a path.
2. **Likelihood Reward:** $r_i^{\text{prob}}(s, a) = \pi_i(s, a)$. This reward uses the direct probability, ensuring non-negativity ($r_i^{\text{prob}} \geq 0$). This positivity is needed for social welfare functions involving products or ratios, such as Nash Welfare.

[Raf+24]: Rafailov et al. (2024), 'From r to Q^*: Your Language Model is Secretly a Q-Function'

We denote by $\mathscr{C}$ the set of all possible complete paths (sequences ending in $\langle \text{eos} \rangle$) from s_0.

In practice, we assume at each non-terminal state s, we only consider a finite set of B possible next tokens $A_B(s) \subseteq A$. This set could be chosen by a base language model ranking tokens by probability, effectively defining a limited vocabulary for each generation step. With this finite branching factor B and a maximum sequence length $L_{\max}$, the set $\mathscr{C}$ of complete paths is finite (bounded by $B^{L_{\max}}$).

This sequential token selection process naturally defines a tree structure rooted at s_0. Each edge represents choosing a token from $A_B(s_t)$, and each node represents a partial sequence s_t. To make this concrete, consider a scenario with two agents ($N = 2$) deciding if a community event should feature 'music' or 'food'. If the generation reaches the state s_t = 'Event should feature', the set $A_B(s_t)$ might contain {' music', ' food', ...}. Figure 3.1 provides a conceptual illustration of this generation tree.

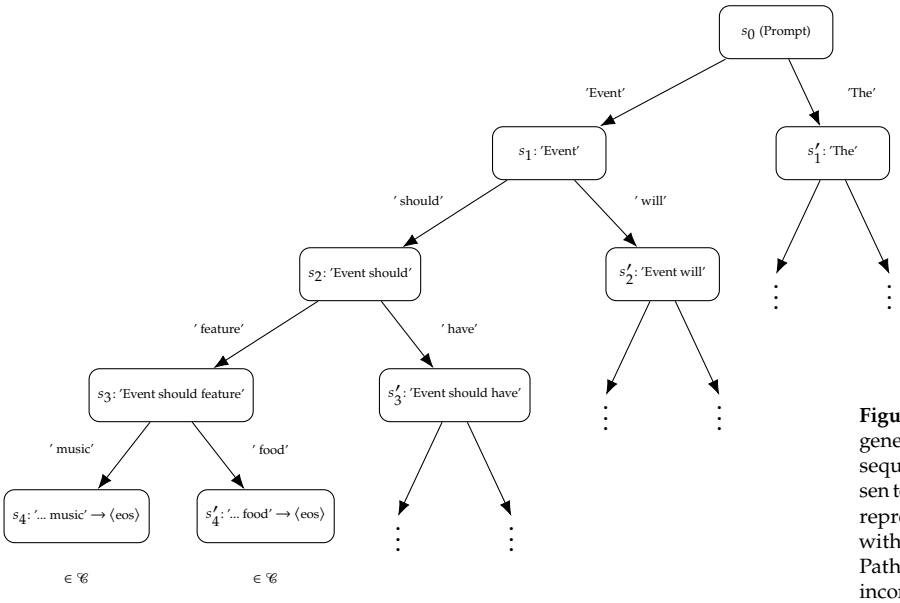


Figure 3.1: Illustration of the token-level generation tree. Each state s_t is a partial sequence. Edges represent actions (chosen tokens). Paths ending in a boxed node represent completed sequences ending with ⟨eos⟩, which belong to the set $\mathcal{C}$. Paths ending in vertical dots represent incomplete branches.

The leaves of this tree, corresponding to paths ending in ⟨eos⟩, constitute the set $\mathcal{C}$. The challenge, particularly given potential agent disagreements at branching points, is to define fair procedures for selecting either a single path (statement) or a distribution over paths in $\mathcal{C}$.

Agent Utilities. Given a completed sequence $X = (a_1, \ldots, a_\ell = \langle\text{eos}\rangle)$ corresponding to states $(s_0, s_1, \ldots, s_\ell)$, we define two corresponding utility functions for each agent i, derived from the respective reward types:

1. **Additive Log-Utility:** Primarily used for evaluating single paths based on cumulative log-likelihood.

$$\begin{aligned} U_i^{\log}(X) &= \sum_{t=1}^{\ell} r_i^{\log}(s_{t-1}, a_t) = \sum_{t=1}^{\ell} \beta \log \pi_i(s_{t-1}, a_t) \\ &= \beta \log \left(\prod_{t=1}^{\ell} \pi_i(s_{t-1}, a_t) \right) \end{aligned}$$

2. **Multiplicative Probability Utility:** Primarily used for evaluating distributions via expected utility, forming the basis for Nash Welfare and Proportional Fairness calculations.

$$U_i^{\text{prob}}(X) = \prod_{t=1}^{\ell} r_i^{\text{prob}}(s_{t-1}, a_t) = \prod_{t=1}^{\ell} \pi_i(s_{t-1}, a_t) = P_i(X)$$

   This represents the joint probability (or likelihood product) of sequence X under agent i's policy, ensuring $U_i^{\text{prob}}(X) \geq 0$.

These are related by $U_i^{\log}(X) = \beta \log U_i^{\text{prob}}(X)$, with $U_i^{\text{prob}}(X) > 0$.

Single Paths vs. Distributions. We consider two types of outcomes:

1. A single deterministic path $X \in \mathcal{C}$. Fairness for single paths can be assessed using criteria like Egalitarian Welfare applied to the additive log-utilities $U_i^{\log}(X)$.
2. A distribution $p \in \Delta(\mathcal{C})$ over paths (a lottery). Fairness for lotteries relies on criteria like Proportional Fairness and the Core, which operate on expected utilities. For these, the probability-based utility is necessary:

$$U_i^{\text{prob}}(p) = \mathbb{E}_{X\sim p}[U_i^{\text{prob}}(X)] = \sum_{X\in\mathcal{C}} p(X)U_i^{\text{prob}}(X).$$

This expected utility $U_i^{\text{prob}}(p)$ is guaranteed to be non-negative, satisfying requirements of downstream fairness measures.

The Fairness of a Distribution. We analyze the fairness of lotteries $p \in \Delta(\mathcal{C})$ using criteria defined over the non-negative expected utilities $U_i^{\text{prob}}(p)$. We employ *proportional fairness*, which is equivalent to maximizing the Nash social welfare (NW) function [Eba+24].

[Eba+24]: Ebadian et al. (2024), 'Optimized distortion and proportional fairness in voting'

The Nash welfare function requires positive inputs, making the probability-based expected utilities suitable:

$$\text{NW}(p) = \prod_{i=1}^{n} U_i^{\text{prob}}(p)$$

Maximizing this product is equivalent to maximizing its logarithm, $\log(\text{NW}(p)) = \sum_{i=1}^{n} \log(U_i^{\text{prob}}(p))$. The use of $U_i^{\text{prob}}(p)$ ensures the arguments of the product and logarithm are non-negative (and typically positive under mild assumptions), unlike expected log-utilities, which are non-positive.

Proportional fairness (PF) itself is defined via ratios of these positive utilities:

$$\text{PF}(p) = \max_{p'\in\Delta(\mathcal{C})} \frac{1}{n} \sum_{i\in N} \frac{U_i^{\text{prob}}(p')}{U_i^{\text{prob}}(p)}$$

As mentioned above, a distribution p is proportionally fair if $\text{PF}(p) = 1$. Maximizing NW is known to yield a proportionally fair distribution [Eba+24].

[Eba+24]: Ebadian et al. (2024), 'Optimized distortion and proportional fairness in voting'

We can also evaluate the fairness of a distribution using the concept of the *core* which originates from cooperative game theory. The core is a strong notion of stability [Sha71]. In the context of collective decision-making, including voting and resource allocation, an outcome in the core signifies that no subgroup (coalition) of agents can unilaterally break away and achieve a different outcome that all members of that subgroup strictly prefer (or where some prefer and none are worse off), considering their proportional influence [ABM19]. This resistance to coalitional deviation is highly desirable for consensus statements, as it suggests the outcome (a lottery in this case) is robust and less likely to be rejected by dissatisfied subgroups who feel they could engineer a universally better distribution over alternatives for themselves using their proportional share of probability mass. *In essence, a lottery in the core represents an outcome that "everyone would agree to" making it a highly appealing notion in the context of consensus generation.* Maximizing Nash

[Sha71]: Shapley (1971), 'Cores of convex games'

[ABM19]: Aziz et al. (2019), 'Fair mixing: the case of dichotomous preferences'

Welfare of expected utility provides a computationally grounded way to find a distribution within the core [FMS18; Eba+24].

[FMS18]: Fain et al. (2018), 'Fair allocation of indivisible public goods'
[Eba+24]: Ebadian et al. (2024), 'Optimized distortion and proportional fairness in voting'

We define the core using probability-based expected utilities $U_i^{\text{prob}}(p)$ since they are non-negative and thus are compatible with Nash welfare optimization.

Definition 3.3.1 (α-Core) *For $\alpha \geq 1$, a distribution $p \in \Delta(\mathcal{C})$ is in the α-core if there is no coalition $S \subseteq N$ and alternative distribution p' such that*

$$\frac{|S|}{|N|} \cdot U_i^{prob}(p') \geq \alpha \cdot U_i^{prob}(p), \quad \forall i \in S,$$

with strict inequality for at least one agent $i \in S$. When $\alpha = 1$, this is also simply referred to as the core.

A key result is that a distribution maximizing Nash Welfare (using U_i^{prob}) is guaranteed to be in the 1-core [ABM19; FMS18]. This provides a strong stability guarantee for our lottery-based approach.

[ABM19]: Aziz et al. (2019), 'Fair mixing: the case of dichotomous preferences'
[FMS18]: Fain et al. (2018), 'Fair allocation of indivisible public goods'

The Fairness of a Path When selecting a single path $X \in \mathcal{C}$, we focus on fairness criteria addressing the distribution of outcomes across agents for that specific path. In particular, we attempt to maximize the sum or the minimum utility. For this criterion, the additive structure of the log-utility $U_i^{\log}(X)$ is natural. For instance, the *egalitarian welfare* of path X is defined using log-utilities as:

$$\text{EW}^{\log}(X) = \min_{i \in N} U_i^{\log}(X) = \min_{i \in N} \sum_{t=1}^{\ell} \beta \log \pi_i(s_{t-1}, a_t) \tag{3.1}$$

Maximizing $\text{EW}^{\log}(X)$ seeks a path whose cumulative log-likelihood is highest for the agent who values it least.

When selecting a single outcome, the egalitarian welfare criterion, $\min_{i \in N} U_i(X)$, draws inspiration from Rawls' theory of justice, particularly the "maximin" principle, which advocates for maximizing the well-being of the least advantaged member of society [Raw71]. In the context of generating a single consensus statement from diverse opinions, maximizing the minimum utility (or minimizing the maximum dissatisfaction) across all agents addresses the risk of marginalizing minority viewpoints or producing a statement that, while potentially pleasing a majority, is deeply unacceptable to some participants. By focusing on the agent who benefits least, the egalitarian objective promotes broadly acceptable outcomes, fostering inclusivity and reducing the potential for alienation. *Therefore, ensuring a minimum level of satisfaction for the least-advantaged participant directly supports the fundamental requirement that a consensus statement ultimately be acceptable to all involved parties.*

[Raw71]: Rawls (1971), 'An egalitarian theory of justice'

3.4 Stochastic Policies and Ex-Ante Fairness

Having established the token-level MDP and fairness criteria, our goal is to define a *stochastic generation policy* Π^* that produces a distribution

p_{Π^*} over complete consensus statements $\mathcal{C}$ satisfying strong ex-ante fairness properties. Specifically, we aim for the generated distribution to be **proportionally fair** and reside in the **1-core**, ensuring stability against coalitional deviations.

As discussed previously, a distribution $p^* \in \Delta(\mathcal{C})$ that maximizes the Nash Welfare (NW) product, $\prod_{i=1}^{n} U_i^{\text{prob}}(p)$, achieves both proportional fairness and 1-core membership [ABM19; FMS18; Eba+24]. Therefore, our strategy is twofold: first, identify this optimal target distribution p^*; second, derive a sequential policy Π^* that, when executed step-by-step during generation, yields exactly p^*.

[ABM19]: Aziz et al. (2019), 'Fair mixing: the case of dichotomous preferences'
[FMS18]: Fain et al. (2018), 'Fair allocation of indivisible public goods'
[Eba+24]: Ebadian et al. (2024), 'Optimized distortion and proportional fairness in voting'

A primary challenge arises from the potentially enormous size of the set of all possible statements $\mathcal{C}$, which can grow exponentially with the maximum sequence length $L_{\max}$ and branching factor B. Directly computing p^* over the full space $\mathcal{C}$ is often computationally intractable.

3.4.1 Computational Tractability via Token Chunking

To address the computational burden, we introduce a *chunking* strategy. This approach groups sequences of tokens into larger units, effectively reducing the depth and branching factor of the decision tree.

Definition 3.4.1 (Token Chunking) *A chunking strategy $\mathcal{K}$ partitions the token sequence into contiguous segments (chunks) $\{k_1, k_2, \ldots, k_m\}$. Each chunk k_j consists of one or more tokens, and their concatenation forms the complete statement. Actions in the MDP now correspond to selecting entire chunks.*

We employ a simple fixed-size chunking strategy where each chunk has size c (a hyperparameter). This reduces the effective path length from $L_{\max}$ tokens to $\lceil L_{\max}/c \rceil$ chunks. The search space is thus restricted to $\mathcal{C}_{\mathcal{K}}$, the set of complete paths constructible using these chunks.

This chunking significantly improves computational feasibility. However, it introduces an approximation: the optimal proportionally fair lottery over the full space $\mathcal{C}$ might involve paths not representable in the chunked space $\mathcal{C}_{\mathcal{K}}$. Our method finds the lottery $p^* \in \Delta(\mathcal{C}_{\mathcal{K}})$ that maximizes Nash Welfare *relative to this restricted space*. Consequently, the fairness guarantees (proportional fairness, 1-core membership) hold within $\mathcal{C}_{\mathcal{K}}$, but the expected utilities achieved might be suboptimal compared to what was possible in the unchunked space $\mathcal{C}$. The choice of chunk size c thus involves a trade-off between tractability and the potential optimality gap.

3.4.2 Deriving the Ex-Ante Fair Stochastic Policy

Given the target distribution $p^* \in \Delta(\mathcal{C}_{\mathcal{K}})$ that maximizes Nash Welfare over the chunked space, we now derive the stochastic policy Π^* that generates this distribution. Executing Π^* involves sequentially sampling the next chunk based on probabilities derived from p^*.

To define Π^*, we introduce some notation. For any state (prefix) s in the chunked generation tree:

- Let $\mathscr{C}_{\mathscr{K}}(s) \subseteq \mathscr{C}_{\mathscr{K}}$ be the set of all complete paths (leaves) in the chunked space that pass through state s.
- Let $\mathscr{C}_{\mathscr{K}}(s,k) \subseteq \mathscr{C}_{\mathscr{K}}(s)$ be the subset of paths in $\mathscr{C}_{\mathscr{K}}(s)$ where the next action (chunk) taken from state s is k. Note that $\mathscr{C}_{\mathscr{K}}(s,k) = \mathscr{C}_{\mathscr{K}}(s\|k)$, where $s\|k$ is the state reached after taking chunk k.
- For any subset of leaves $L \subseteq \mathscr{C}_{\mathscr{K}}$, let $P^*(L) = \sum_{X \in L} p^*(X)$ be the total probability mass assigned by the optimal lottery p^* to the leaves in L.

Note that $P^*(\mathscr{C}_{\mathscr{K}}(s_0)) = P^*(\mathscr{C}_{\mathscr{K}}) = 1$, where s_0 is the initial empty state.

With this notation, we can define the policy Π^* at any given state s.

Definition 3.4.2 (Stochastic Policy Induced by Lottery p^*) *Let p^* be a distribution over the leaf nodes $\mathscr{C}_{\mathscr{K}}$. The induced stochastic policy Π^* at a non-terminal state s assigns the probability of taking the next action (chunk) k as:*

$$\Pi^*(s,k) = \begin{cases} \frac{P^*(\mathscr{C}_{\mathscr{K}}(s,k))}{P^*(\mathscr{C}_{\mathscr{K}}(s))} & \text{if } P^*(\mathscr{C}_{\mathscr{K}}(s)) > 0 \\ 0 & \text{if } P^*(\mathscr{C}_{\mathscr{K}}(s)) = 0 \end{cases} \tag{3.2}$$

This represents the conditional probability, according to the target distribution p^*, of selecting chunk k next, given that the generation process has reached state s. If state s has zero probability of being reached under p^* (i.e., $P^*(\mathscr{C}_{\mathscr{K}}(s)) = 0$), then the probability of taking any action from s is also zero.

The relationship between the target distribution p^* over leaves and the calculation of the policy Π^* at an internal node s is illustrated in Figure 3.2.

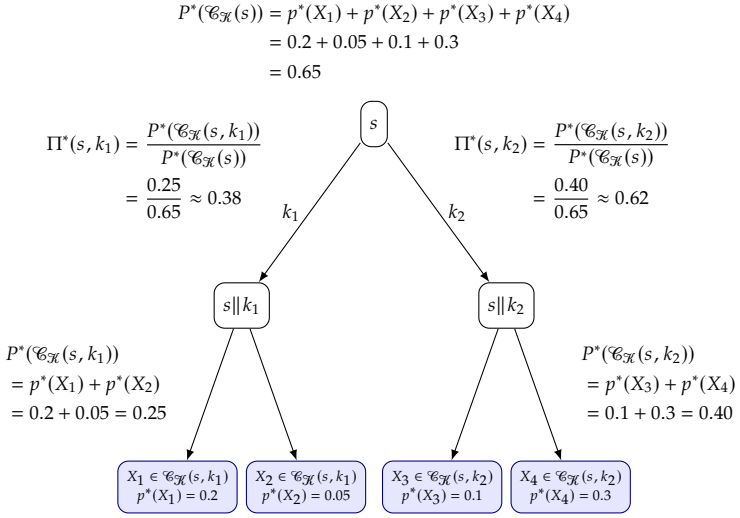


Figure 3.2: Illustration of the induced stochastic policy Π^* at state s. The optimal lottery p^* assigns probabilities to the leaf nodes (complete paths). The probability $P^*(\mathscr{C}_{\mathscr{K}}(s))$ is the sum of $p^*(X)$ for all leaves reachable from s. The probability $P^*(\mathscr{C}_{\mathscr{K}}(s,k))$ is the sum for leaves reachable via action k. The policy $\Pi^*(s,k)$ is the conditional probability of taking action k.

3.4.3 Properties of the Induced Policy

We now establish that executing this policy Π^* from the initial state s_0 indeed generates the target distribution p^* over the leaves of the chunked tree $\mathscr{C}_{\mathscr{K}}$.

Theorem 3.4.1 (Equivalence of Policy-Induced Distribution and Target Lottery) *Let p_{Π^*} be the distribution over $\mathscr{C}_{\mathcal{K}}$ generated by executing the policy Π^* (defined in Def. 3.4.2) from the initial state s_0. Then $p_{\Pi^*} = p^*$.*

Proof. We prove by induction on the depth of state s in the chunked tree that the probability of reaching state s under policy Π^*, denoted $P_{\Pi^*}(s)$, is equal to $P^*(\mathscr{C}_{\mathcal{K}}(s))$, the total mass assigned by p^* to leaves passing through s.

Base Case (Depth 0): The initial state is s_0. $P_{\Pi^*}(s_0) = 1$ by definition. Also, $\mathscr{C}_{\mathcal{K}}(s_0) = \mathscr{C}_{\mathcal{K}}$ (all paths pass through the start state), and $P^*(\mathscr{C}_{\mathcal{K}}(s_0)) = \sum_{X \in \mathscr{C}_{\mathcal{K}}} p^*(X) = 1$ since p^* is a probability distribution. Thus, $P_{\Pi^*}(s_0) = P^*(\mathscr{C}_{\mathcal{K}}(s_0))$.

Inductive Hypothesis (IH): Assume that for all states s at depth d, $P_{\Pi^*}(s) = P^*(\mathscr{C}_{\mathcal{K}}(s))$.

Inductive Step: Consider an arbitrary state s' at depth $d+1$. State s' must be reached from a unique predecessor state s at depth d by taking a specific action (chunk) k, such that $s' = s \| k$. The probability of reaching s' under Π^* is:

$$\begin{aligned} P_{\Pi^*}(s') &= P_{\Pi^*}(s) \cdot \Pi^*(s, k) \\ &= P^*(\mathscr{C}_{\mathcal{K}}(s)) \cdot \Pi^*(s, k) \quad \text{(by IH)} \end{aligned}$$

If $P^*(\mathscr{C}_{\mathcal{K}}(s)) = 0$, then $P_{\Pi^*}(s) = 0$, implying $P_{\Pi^*}(s') = 0$. Also, if $P^*(\mathscr{C}_{\mathcal{K}}(s)) = 0$, then $P^*(\mathscr{C}_{\mathcal{K}}(s, k)) = 0$ since $\mathscr{C}_{\mathcal{K}}(s, k) \subseteq \mathscr{C}_{\mathcal{K}}(s)$. Since $s' = s \| k$, $\mathscr{C}_{\mathcal{K}}(s') = \mathscr{C}_{\mathcal{K}}(s, k)$, so $P^*(\mathscr{C}_{\mathcal{K}}(s')) = 0$. Thus, $P_{\Pi^*}(s') = P^*(\mathscr{C}_{\mathcal{K}}(s')) = 0$.

If $P^*(\mathscr{C}_{\mathcal{K}}(s)) > 0$, we use the definition of $\Pi^*(s, k)$:

$$\begin{aligned} P_{\Pi^*}(s') &= P^*(\mathscr{C}_{\mathcal{K}}(s)) \cdot \frac{P^*(\mathscr{C}_{\mathcal{K}}(s, k))}{P^*(\mathscr{C}_{\mathcal{K}}(s))} \\ &= P^*(\mathscr{C}_{\mathcal{K}}(s, k)) \end{aligned}$$

Since $s' = s \| k$, we have $\mathscr{C}_{\mathcal{K}}(s') = \mathscr{C}_{\mathcal{K}}(s, k)$. Therefore,

$$P_{\Pi^*}(s') = P^*(\mathscr{C}_{\mathcal{K}}(s'))$$

This completes the inductive step.

Conclusion: The induction holds for all states s. Now, consider any leaf node $X \in \mathscr{C}_{\mathcal{K}}$. A leaf node is a state at the maximum depth. The set of paths passing through leaf X is just the singleton set $\{X\}$, so $\mathscr{C}_{\mathcal{K}}(X) = \{X\}$. Applying our proven result for state $s = X$:

$$P_{\Pi^*}(X) = P^*(\mathscr{C}_{\mathcal{K}}(X)) = P^*(\{X\}) = p^*(X)$$

Since this holds for all $X \in \mathscr{C}_{\mathcal{K}}$, the distribution p_{Π^*} induced by policy Π^* is identical to the target distribution p^*. □

This equivalence directly leads to the desired ex-ante fairness guarantee for the policy Π^*.

Corollary 3.4.2 (Core Membership of Stochastic Policy) *Let p^* be a distribution over $\mathscr{C}_{\mathcal{K}}$ that maximizes Nash Welfare (and is therefore in the 1-core relative to $\mathscr{C}_{\mathcal{K}}$). Let Π^* be the stochastic policy derived from p^* according to Definition 3.4.2. Then the distribution p_{Π^*} generated by executing Π^* is in the 1-core relative to $\mathscr{C}_{\mathcal{K}}$.*

Proof. By Theorem 3.4.1, the distribution generated by policy Π^* is $p_{\Pi^*} = p^*$. Since p^* was chosen to maximize Nash Welfare over $\mathscr{C}_{\mathcal{K}}$, it is in the 1-core relative to this set. Therefore, p_{Π^*} is also in the 1-core relative to $\mathscr{C}_{\mathcal{K}}$. □

This result confirms that our procedure—finding the MNW distribution p^* over the chunked space $\mathscr{C}_{\mathcal{K}}$ and then executing the derived policy Π^*—yields a generation process whose resulting distribution satisfies the 1-core stability property within that space.

3.4.4 A Note on Computation and Approximation

While chunking makes the state space $\mathscr{C}_{\mathcal{K}}$ more manageable than $\mathscr{C}$, computing the exact optimal distribution p^* that maximizes Nash Welfare over $\mathscr{C}_{\mathcal{K}}$ can still be computationally demanding, requiring analysis of the entire chunked tree. Yet, the definition and properties of the induced policy Π^* are still theoretically valuable as they establish the existence of a stochastic policy in the core.

This characterization serves as a foundation for future work on practical approximation algorithms. Such algorithms could aim to find policies $\hat{\Pi}^*$ that approximate Π^* without needing to compute p^* explicitly or explore the full chunked tree. For instance, one could investigate whether sampling-based methods, such as variants of Monte Carlo Tree Search adapted for the Nash Welfare objective [HMS21], could approximate the target policy Π^* and potentially converge towards the optimal distribution p^*. The goal would be to achieve near-optimal Nash welfare efficiently. Furthermore, established theoretical results connect the degree of approximation to optimal Nash welfare with guarantees related to α-core membership [Eba+24]. This suggests that successful approximation algorithms could yield generation policies with quantifiable α-core stability guarantees, providing a practical path towards implementing these ex-ante fair stochastic policies.

[HMS21]: Hossain et al. (2021), 'Fair algorithms for multi-agent multi-armed bandits'

[Eba+24]: Ebadian et al. (2024), 'Optimized distortion and proportional fairness in voting'

3.5 Generating a Single Consensus Statement

While our proportionally fair stochastic policy Π^* offers strong ex-ante fairness guarantees, many practical applications require selecting a single consensus statement. When generating such a statement, our objective shifts from finding a fair distribution to identifying the single path that best represents all agents' preferences.

Algorithm 2 Finite Lookahead Egalitarian Welfare Maximization

Require: Set of agents N, Lookahead depth d, Branching factor B, Max length $L_{\max}$

Initialize current state s_0 to the empty sequence; $t \leftarrow 0$
Initialize generated path $X_{fl} \leftarrow (s_0)$
while $t < L_{\max}$ and s_t is not terminal **do**
    Let X_{prefix} be the path corresponding to s_t.
    Let $\mathcal{P}_d(s_t)$ be the set of all paths $P = (a_1, \ldots, a_k)$ starting from s_t such that $k \leq d$ and $X_{prefix} \| P$ does not exceed length $L_{\max}$.
    Find a path $P^* = (a_1^*, \ldots, a_{k^*}^*) \in \mathcal{P}_d(s_t)$ that maximizes the lookahead objective:

$$\max_{P \in \mathcal{P}_d(s_t)} \min_{i \in N} U_i^{\log}(X_{prefix} \| P)$$

    if no path P^* found (e.g., s_t is terminal) **then**
        Break
    end if
    Take the first action $a^* \leftarrow a_1^*$.
    Update state: $s_{t+1} \leftarrow T(s_t, a^*)$
    Append a^* to the sequence represented by X_{fl}.
    $t \leftarrow t + 1$
end while
return Complete path X_{fl}

3.5.1 The Rawlsian Path

Given the token tree with leaf nodes $\mathcal{C}$ (potentially derived from chunking, $\mathcal{C}_{\mathcal{K}}$), we aim to find a single path $X^* \in \mathcal{C}$ that maximizes egalitarian welfare as defined in Equation 3.1. Due to the size of the token tree, exhaustive search for X^* may be intractable. We propose approximate algorithms to find high-quality paths.

First, we propose a finite-lookahead algorithm. At each timestep, this algorithm examines all paths extending d steps into the future from the current state. It selects the initial action belonging to the path that yields the highest egalitarian welfare within this d-step horizon. This approach can mitigate the potential for hedging inherent in greedy search. Specifically, when no single immediate token garners broad agreement (i.e., results in high egalitarian welfare), a greedy method might select less informative tokens that avoid commitment. In contrast, a lookahead can identify longer sequences that, despite potentially controversial initial steps, lead to states with higher overall welfare, perhaps by expressing a concept with suitable qualifications..

Second, we consider beam search. This is a widely used heuristic search algorithm in natural language processing with demonstrated empirical effectiveness.

Finite Lookahead Search

The finite lookahead algorithm operates with a rolling horizon. At each step t (current state s_t), it explores all possible paths P of length up to d originating from s_t. For each such path P, it evaluates the egalitarian welfare of the sequence formed by concatenating the path generated so far (X_{prefix}) with P. It then chooses the *first action* a^* of the path P^* that

Algorithm 3 Egalitarian Welfare Beam Search

Require: Set of agents N, Beam width w, Branching factor B, Max length $L_{\max}$

Initialize beam $\mathcal{B}_0 = \{(s_0, \text{path } s_0)\}$ with the empty sequence path
for $t = 0$ to $L_{\max} - 1$ **do**
  $\mathcal{C}_{t+1} \leftarrow \emptyset$ ▷ Candidate set for next beam
  for each path X_{path} represented by state sequence $(s_0, \ldots, s_t)$ in $\mathcal{B}_t$ **do**
    if s_t is not terminal **then**
      Consider B possible next actions $A_B(s_t)$ from state s_t
      for each action $a \in A_B(s_t)$ **do**
        $s_{t+1} \leftarrow T(s_t, a)$
        $X_{new_path} \leftarrow X_{path} \| a$ ▷ Append action to path sequence
        Add X_{new_path} (represented by its state sequence) to $\mathcal{C}_{t+1}$
      end for
    else
      Add X_{path} (already terminal) to $\mathcal{C}_{t+1}$
    end if
  end for
  For each path $X \in \mathcal{C}_{t+1}$, compute its score $f(X) = \min_{i \in N} U_i^{\log}(X)$.
  $\mathcal{B}_{t+1} \leftarrow$ top w paths from $\mathcal{C}_{t+1}$ according to score $f(X)$.
end for
return Path from final beam $\mathcal{B}_{L_{\max}}$ with the highest score $f(X)$.

maximizes this lookahead evaluation, transitions to state $s_{t+1} = T(s_t, a^*)$, and repeats.

Beam Search

Beam search is a widely used heuristic search algorithm that offers a practical compromise between greedy search and exhaustive exploration, particularly effective in sequence generation tasks like machine translation and text generation [MJ09; Raf+24]. Instead of pursuing only the single best option at each step (like greedy search) or all options (like exhaustive search), beam search maintains a fixed number, w (the beam width), of the most promising partial paths (hypotheses) at each depth t. It expands all paths currently in the beam, generates their potential successor tokens/paths, evaluates all resulting candidates using the chosen objective function (here, egalitarian welfare), and retains only the top w scoring paths for the next step. This allows it to explore a more diverse set of sequences than greedy search, reducing the risk of getting stuck in suboptimal paths due to poor early choices, while remaining computationally tractable compared to exploring the entire exponentially large search space. The final result is the highest-scoring complete path found among the beams at the maximum length or upon reaching terminal states.

[MJ09]: Martin et al. (2009), *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*
[Raf+24]: Rafailov et al. (2024), 'From r to Q^*: Your Language Model is Secretly a Q-Function'

Reference policy prompt	User policy prompt	Sequence	Z-Score
User food profile: empty	User food profile: vegetarian	I am having chicken/tofu enchiladas tonight. Then I am going to meet up with some friends.	2.69
User location profile: empty	User location profile: lives in a cold climate	I 'm going to the beach/mountains this weekend to surf/ski . I need to buy some new clothes .	1.78, 3.31
User time profile: empty	User time profile: morning	I am about to eat some food . I am going to have spaghetti/pancakes . I will use my phone to order it .	4.26
User opinion: empty	User opinion: Favors stricter gun control laws.	Implement ing background checks that are less/more strict for gun purchases is essential . Also , my favorite color is orange .	2.09

Table 3.1: Credit assignment results for `Llama 3.1 8B Instruction-Tuned`. Darker green indicates larger Z-score. Z-score column is for altered tokens. Alterations are represented by "<misaligned>/<aligned>".

3.6 Experiments

We conduct experiments to validate key aspects of our proposed framework and evaluate its practical performance. First, we investigate whether agent policies derived from prompting language models can perform credit-assignment to provide meaningful token-level rewards, which is a prerequisite for our search algorithms. Second, we assess the performance of the finite lookahead and beam search methods in generating single consensus statements, evaluating their ability to optimize egalitarian welfare compared to baseline approaches.

3.6.1 Evaluating Prompt-Based Credit Assignment

Our framework utilizes agent policies π_i, instantiated by prompting a base LLM with agent-specific information, to define token-level rewards $r_i(s, a) = \beta \log \pi_i(s, a)$. Token-level search algorithms guided by such rewards, like those we propose, rely on the underlying policies exhibiting localized credit assignment—meaning the policy likelihoods $\pi_i(s, a)$ should primarily change at tokens relevant to the specific profile information. This property has been observed in models trained with Direct Preference Optimization (DPO) [Raf+24]. To ensure our approach is well-founded, we empirically validate that this necessary property also holds for policies derived from simple system prompting of instruction-tuned models.

[Raf+24]: Rafailov et al. (2024), 'From r to Q^*: Your Language Model is Secretly a Q-Function'

Our experimental setup, inspired by the analysis in Rafailov et al. [Raf+24] but adapted for prompting, involved controlled experiments using `Llama`

[Raf+24]: Rafailov et al. (2024), 'From r to Q^*: Your Language Model is Secretly a Q-Function'

`3.1 8B Instruction-Tuned` and `Gemma 2 9b Instruction-Tuned`[1]. For each test case, we compared token log-probabilities under a specific profile prompt versus a baseline (empty profile) prompt. This comparison was performed for two nearly identical input sequences: one (X_1) containing a concept conflicting with the profile, and another (X_2) containing an aligned concept. To quantify the localized effect of the profile change between these sequences, we calculated the z-score for each token position j. This z-score was based on the absolute change in the token's (profile - baseline) log-probability difference between sequence X_1 and X_2, relative to the mean and standard deviation of these changes across the sequence. Using the z-score helps normalize the changes, making comparisons meaningful across tokens that might have different inherent variability in log-probability differences; it specifically highlights tokens where the profile prompt induced a statistically significant shift in preference when the aligned/misaligned concept was swapped, pinpointing localized credit assignment. A high z-score thus highlights tokens where the profile's influence shifted substantially (e.g., from positive to negative) when the input changed from conflicting to aligning.

1: Results for Gemma 2 9b Instruction-Tuned were similar and are included in Table B.1 in Appendix B.

As hypothesized, the analysis revealed that the largest changes (highest z-scores) were concentrated at the key tokens varied between inputs (e.g., "chicken/tofu" for a vegetarian profile), as shown in Table 3.1. This confirms that the localized sensitivity or credit assignment property can be induced via system prompting in instruction-tuned models, not just via DPO training. This finding validates the foundational premise that the log-likelihoods $r_i(s, a) = \beta \log \pi_i(s, a)$ from our prompted policies provide meaningful, localized signals suitable for guiding the token-level search algorithms.

3.6.2 Consensus Generation

To evaluate the practical effectiveness of different approaches for generating a single consensus statement, we conducted experiments comparing our proposed search algorithms against several baselines. The primary goal was to assess how well each method optimizes the egalitarian welfare (EW) objective, operationalized through a perplexity-based metric reflecting the worst-case agent alignment.

Experimental Setup.

Scenarios: We utilized scenarios from the Habermas Machine dataset [Tes+24]. To select distinct experimental settings, we embedded the scenario descriptions using `BAAI/bge-large-en-v1.5` [Xia+23] and performed k-means clustering with $k = 3$, selecting representative scenarios from each cluster (Scenarios 1, 2, and 3). The Issues from scenarios 1, 2, and 3 are stated in the table captions of Table B.2, Table B.3, and Table B.4, respectively.

[Tes+24]: Tessler et al. (2024), 'AI can help humans find common ground in democratic deliberation'

[Xia+23]: Xiao et al. (2023), *C-Pack: Packaged Resources To Advance General Chinese Embedding*

Agents and Policies: For each scenario, we used the corresponding agent opinions from the dataset. Agent policies π_i were instantiated by prompting the `Llama 3.1 8B Instruct` model [Gra+24] with the specific issue and agent i's opinion, instructing it to generate text aligned with that viewpoint (full prompt in Section B.1 of Appendix B); the

[Gra+24]: Grattafiori et al. (2024), 'The llama 3 herd of models'

resulting likelihoods $\pi_i(s, a)$ represent agent i's preferences. The agent opinions are shown in Table B.2, Table B.3, and Table B.4 in Appendix B.

Base Generation Model: Unless otherwise specified, consensus statements were generated using `Llama 3.1 8B Instruct`, prompted with the issue and all agent opinions (full prompt Section B.1 of Appendix B).

Evaluation Metric - Egalitarian Perplexity (EPPL): We define a metric to capture how well a generated consensus statement X aligns with the least satisfied agent. For each agent i, we calculate their specific perplexity $PPL_i(X)$ by prompting the evaluation model (`Llama 3.1 8B Instruct`) with the issue and agent i's opinion, asking it to generate a statement perfectly reflecting that opinion. We then compute the average log-likelihood of the actual consensus statement $X = (a_1, \ldots, a_L)$ conditioned on this agent-specific prompt:

$$\bar{L}_i(X) = \frac{1}{L} \sum_{t=1}^{L} \log \pi_i(s_{t-1}, a_t | \text{prompt}_i).$$

The agent-specific perplexity is $PPL_i(X) = \exp(-\bar{L}_i(X))$. The final Egalitarian Perplexity for statement X is the maximum perplexity across all agents: $EPPL(X) = \max_{i \in N} PPL_i(X)$. Lower EPPL values indicate better egalitarian welfare.

Seeds: We report the average and standard deviation of the EPPL metric over 3 random seeds for each method and scenario combination.

Methods Compared:

We compared the following methods:

1. **Finite Lookahead:** Our proposed algorithm (Algorithm 2) using the base generation model with lookahead depth $d = 4$ and branching factor $B = 4$. A post-processing "brush up" step was applied using `Llama 3.1 70B Instruct` (prompt in Section B.1 of Appendix B).
2. **Beam Search:** Our proposed algorithm (Algorithm 3) using the base generation model with a beam width of $w = 4$, maintaining 4 candidate paths pruned based on partial EPPL scores.
3. **Best-of-N:** We generated $N = 4$ candidate sequences by sampling from the base model and selecting the one with the lowest EPPL score.
4. **Prompted Habermas Machine**[2]**:** An adaptation by the original authors of the baseline [Tes+24], involving 1 round of critique with 4 candidates; we altered it by generating critiques with the base model (`Llama 3.1 8B Instruct`) conditioned on agent opinions (instead of having the human that wrote the original opinion write the critique), before generating the final statement.
5. **Habermas Machine:** The consensus statement provided in the original Habermas Machine study, generated using Chinchilla-70B.

2: As implemented in https://github.com/google-deepmind/habermas_machine

Results

The results, summarized in Table 3.2 and visualized in Figure 3.3, demonstrate performance differences based on the EPPL metric (lower is

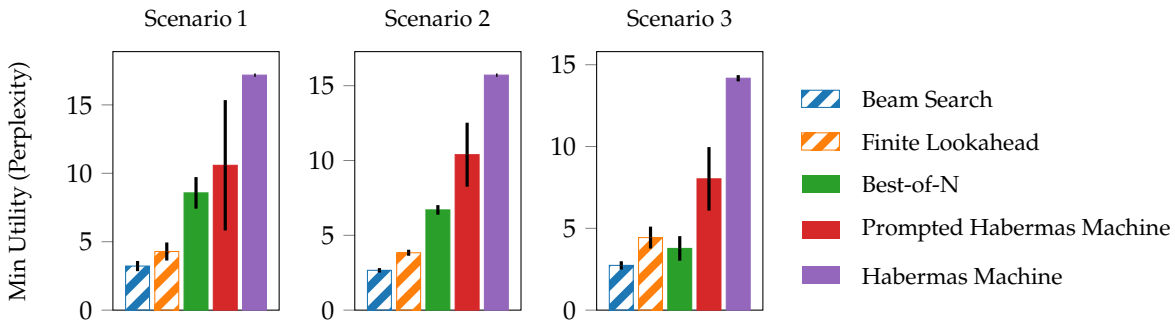


Figure 3.3: Per-scenario egalitarian welfare (perplexity). Lower values indicate better minimum agent utility. Striped bars indicate that the method uses search over the token-level MDP. Results are also shown in Table 3.2.

better). **Beam Search consistently achieved the lowest EPPL** across all scenarios (overall average EPPL: 2.87), indicating the best alignment with the least satisfied agent. **Finite Lookahead also performed well** (overall average EPPL: 4.18), on average yielding lower EPPL than sampling-based and baseline methods. Both search-based methods outperformed **Best-of-N** (6.35) and the **Prompted Habermas Machine** (9.67). The **Habermas Machine** baseline, despite being generated by a larger model, resulted in the highest EPPL (15.69), although this is likely due to the fact that the statement was generated by another model (Chinchilla 70B).

These findings suggest that explicitly searching the token-level MDP guided by the Egalitarian Perplexity objective, as done in Beam Search and Finite Lookahead, is an effective strategy for generating consensus statements that better satisfy the minimum agent alignment requirement compared to sampling-based methods (Best-of-N) or iterative refinement approaches like the Prompted Habermas Machine. The strong performance of Beam Search, in particular, highlights the benefit of exploring a diverse set of promising paths during generation while optimizing egalitarian welfare.

The consensus statements produced by each method are shown for the first seed in Table B.2, Table B.3, and Table B.4.

3.7 Discussion

This work introduced a framework for generating consensus statements by modeling the process as a multi-objective, token-level MDP where rewards are derived from agent-specific language model policies. We aimed to bridge the gap between the flexibility of LLM-based text generation and the formal fairness guarantees of social choice theory via the token-level MDP.

Our theoretical contributions center on adapting established fairness concepts to this generative setting. For stochastic outcomes (lotteries over statements), we focused on proportional fairness (PF) and the core. By maximizing the Nash Welfare product over expected probability-based utilities, we identified an optimal lottery p^*. We demonstrated how this lottery induces a stochastic generation policy Π^* (Definition 3.4.2) that inherits the 1-core property (Corollary 3.4.2), guaranteeing ex-ante stability against coalitional deviations. This provides a principled

Method	Scenario 1	Scenario 2	Scenario 3	Overall
Beam Search	**3.22 ± 0.37**	**2.66 ± 0.15**	**2.74 ± 0.25**	**2.87 ± 0.37**
Finite Lookahead	4.29 ± 0.66	3.83 ± 0.20	4.43 ± 0.67	4.18 ± 0.61
Best-of-N	8.57 ± 1.15	6.70 ± 0.32	3.77 ± 0.75	6.35 ± 2.14
Habermas Machine	10.58 ± 4.76	10.39 ± 2.14	8.03 ± 1.94	9.67 ± 3.42
Predefined	17.18 ± 0.00	15.71 ± 0.00	14.18 ± 0.19	15.69 ± 1.23

Table 3.2: Egalitarian Welfare (Perplexity) Across All Scenarios. Lower is better.

approach for fair randomization in consensus generation, although computing the exact optimal lottery remains challenging. Chunking was introduced as a practical heuristic to manage the search space size.

For deterministic outcomes (single statements), we focused on maximizing egalitarian welfare (EW), aiming to find the path X maximizing $\min_i U_i^{\log}(X)$. Recognizing the computational cost of finding the optimal path, we proposed approximation algorithms: finite lookahead search and beam search.

The empirical results provide practical validation for key aspects of our framework. First, the credit assignment experiments (Section 3.6.1) confirmed that prompting LLMs with agent profiles effectively creates policies π_i whose token likelihoods reflect the specified viewpoints. This finding supports the viability of using $r_i(s, a) = \beta \log \pi_i(s, a)$ as meaningful reward signals within the MDP. Second, the consensus generation experiments (Section 3.6.2) compared methods based on Egalitarian Perplexity (EPPL), a measure inversely related to EW. Both beam search and finite lookahead search, guided by the EW objective, consistently outperformed baseline methods like Best-of-N sampling and an adapted Habermas Machine approach across different scenarios using Llama 3.1 models. Beam search yielded the lowest EPPL, suggesting its effectiveness in navigating the trade-off between exploration breadth and objective optimization for finding high-EW single paths. The relatively poor performance of the predefined Habermas Machine baseline might stem from model differences or the specific nature of the EW objective compared to the objective of the Habermas Machine [Tes+24].

[Tes+24]: Tessler et al. (2024), 'AI can help humans find common ground in democratic deliberation'

Overall, the results suggest that formulating consensus generation as a search problem within a token-level MDP, guided by explicit social choice objectives (like EW for single paths), is a promising direction. The search-based methods (beam search, finite lookahead) demonstrated clear advantages over sampling or baseline iterative methods in optimizing for the minimum agent utility, as measured by EPPL. While the theoretical guarantees for PF/Core apply to distributions generated by potentially complex stochastic policies, the practical success of EW-guided search provides concrete methods for generating single fair consensus statements. This work thus contributes both theoretical foundations and practical algorithms for incorporating social choice principles into generative AI for collective decision-making tasks, advancing the AI → SC direction outlined in the thesis introduction (Chapter 1).

Conclusion & Future Work 4

4.1 Future Work 41
4.1.1 Future Work: AI for Social Choice (AI → SC) 41
4.1.2 Future Work: Social Choice for AI (SC → AI) 42
4.2 Concluding Remarks . . . 43

This thesis investigated the intersection of social choice theory and artificial intelligence, focusing on the challenge of managing value pluralism in AI alignment and AI-assisted collective decision-making. We explored two primary directions: applying social choice principles to align AI systems with diverse human preferences (SC → AI) and leveraging AI techniques, particularly large language models, to enhance mechanisms for eliciting and synthesizing these preferences (AI → SC). The core argument presented is that LLMs, when combined with principles from social choice, provide valuable tools for both eliciting rich, individualized preferences through reflective processes and for aggregating these diverse viewpoints into fair collective outputs, thereby advancing methods for both pluralistic AI alignment and broader collective sense-making.

The research presented contributes computational methods addressing key aspects of this argument. Chapter 2 introduced the Interactive-Reflective Dialogue Alignment (IRDA) system. IRDA utilizes LLMs to facilitate reflective dialogues, guiding users to articulate and construct their preferences regarding AI behavior in specific contexts. This approach directly addresses the challenge that preferences, especially in novel domains, are often constructed rather than merely revealed. By generating individualized, LLM-based reward models from these dialogues, IRDA offers a method for capturing preference diversity with high sample efficiency, providing a foundation for explicit, principled aggregation informed by social choice theory, rather than relying on the implicit aggregation common in standard RLHF pipelines. Empirical results demonstrated IRDA's ability to model individual values more accurately than non-reflective baselines, particularly when underlying preferences exhibited significant heterogeneity.

Chapter 3 shifted focus to the aggregation stage, specifically addressing the fair generation of consensus statements from multiple viewpoints. We modeled this task as a token-level Markov Decision Process (MDP), where agent preferences are represented by individual policies, often derived from LLMs conditioned on agent opinions. This framework allows for the direct application of social choice criteria within the generation process itself. We proposed two mechanisms: first, a method to compute a distribution over complete statements maximizing proportional fairness (via Nash Welfare), yielding an ex-ante fair stochastic generation policy residing in the 1-core; second, for generating a single statement, we developed deterministic search algorithms (finite lookahead and beam search) guided by the objective of maximizing egalitarian welfare (the utility of the least satisfied agent). Experiments confirmed that prompting instruction-tuned LLMs can induce localized credit assignment suitable for defining token-level rewards and that the proposed search methods, particularly beam search, effectively optimize egalitarian welfare, generating consensus statements with better worst-case agent alignment compared to baseline approaches.

Together, these chapters illustrate a potential pipeline: Chapter 2 offers a method for eliciting and modeling the nuanced, diverse individual

inputs (preferences or policies) required by the fair synthesis mechanisms developed in Chapter 3. This combination provides a more complete approach to handling value pluralism, moving from individual preference construction through reflective interaction to collective output generation guided by formal fairness principles.

4.1 Future Work

The research presented opens several avenues for future investigation, aiming to refine the proposed methods, deepen the integration between the components, and explore broader applications. We organize these directions based on the primary flow of influence between AI and social choice.

4.1.1 Future Work: AI for Social Choice (AI → SC)

This direction focuses on using AI techniques to improve or enable social choice mechanisms, particularly for handling complex, qualitative inputs.

Grounding Consensus Policies with Richer Elicitation The agent policies (π_i) used in Chapter 3 were derived from relatively simple prompts based on static opinion statements. Future work could use the richer, interactively refined preference models or dialogue histories produced by IRDA (Chapter 2) to generate more nuanced and grounded agent policies for the consensus generation MDP, potentially leading to more faithful consensus outputs. Further, an iterative pipeline, similar to that of the Habermas Machine [Tes+24], where preference construction is interleaved with consensus generation, could lead to more nuanced and appealing consensus statements.

[Tes+24]: Tessler et al. (2024), 'AI can help humans find common ground in democratic deliberation'

More Principled Preference Elicitation While IRDA incorporates active learning via uncertainty sampling, exploring more sophisticated methods for selecting trajectories or queries could improve the preference elicitation. Techniques from active reward learning that adaptively sample queries based on inferred user models could be integrated (e.g., Sadigh et al. [Sad+17]). Furthermore, DPO can be applied to multi-turn conversations (in contrast to RLHF [Raf+24]), which provides an avenue for teaching agents how to iteratively elicit preferences in conversation as is explored by Andukuri et al. [And+24].

[Sad+17]: Sadigh et al. (2017), *Active preference-based learning of reward functions*

[And+24]: Andukuri et al. (2024), 'STaR-GATE: Teaching Language Models to Ask Clarifying Questions'

Dynamic and Deliberative Consensus Mechanisms The current consensus model is static. Future work could explore more dynamic approaches where agent policies might evolve, or where a moderator agent (potentially trained via RL within the MDP framework) actively guides a deliberation process, perhaps by posing questions, suggesting compromises, or highlighting areas of disagreement. Perhaps, the Discourse Quality Index [Ste+03] (which was developed to assess the quality of conversation in deliberative settings) could serve as a reward signal.

[Ste+03]: Steenbergen et al. (2003), 'Measuring political deliberation: A discourse quality index'

Generating Diverse Slates via AI Instead of a single consensus statement, adapting the MDP framework to generate a *slate* of diverse statements that collectively represent the range of opinions according to criteria like proportional representation or coverage could be valuable in contexts where a single point of agreement is neither feasible nor desirable. For large token-level MDPs that cannot be fully explored, this requires answering the difficult question of how to use finite exploration resources to satisfy multiple groups. Ideas surrounding metareasoning [RW91] may provide a start on this problem, but would need to turn an eye to fairness.

[RW91]: Russell et al. (1991), 'Principles of metareasoning'

4.1.2 Future Work: Social Choice for AI (SC → AI)

This direction focuses on applying principles and techniques from social choice theory to address challenges in AI, particularly in aligning AI systems with diverse human values.

Applying Social Choice to Aggregate Individualized Reward Models A primary next step is to directly apply social choice aggregation rules, to the individualized reward models generated by IRDA (Chapter 2). This could involve training a single RL policy based on an aggregated reward signal derived from multiple, distinct reward functions. Addressing how to aggregate potentially conflicting reward signals observed over entire trajectories presents a non-trivial technical challenge. Conitzer et al. [Con+24] provide suggestions, but they have yet to be empirically tested.

[Con+24]: Conitzer et al. (2024), 'Position: social choice should guide AI alignment in dealing with diverse human feedback'

Representative Reward Models and Policies Moving beyond individualized reward models, a more scalable approach might involve learning a representative set of reward models or policies, drawing an analogy to how committees are chosen to represent diverse preferences in various settings. While reward models guide policy learning, their correlation with the resulting policies isn't always direct [ZH20], suggesting that directly learning a "policy committee" for AI personalization, as explored by Ge et al. [Ge+25] and Ivanov et al. [IB24], could offer fairness guarantees that are more closely tied to the outcomes users experience. Concepts from social choice theory, such as proportional representation, could ensure that sufficiently large groups of users with similar preferences are catered to by at least one policy.

[ZH20]: Zhuang et al. (2020), 'Consequences of misaligned AI'

[Ge+25]: Ge et al. (2025), 'Learning Policy Committees for Effective Personalization in MDPs with Diverse Tasks'

[IB24]: Ivanov et al. (2024), 'Personalized reinforcement learning with a budget of policies'

Constitutional AI Informed by Social Choice The combined approach of eliciting diverse values (Chapter 2) and synthesizing them using fair mechanisms (Chapter 3) could be applied to the development and refinement of AI constitutions [Bai+22]. This might involve a process where principles are elicited from diverse stakeholders, modeled individually, and then aggregated into a governing constitution using SC-informed methods, ensuring the aggregation process itself adheres to fairness axioms. A preliminary version of this was tested by Huang et al. [Hua+24], but lacked formal fairness guarantees that social choice could provide.

[Hua+24]: Huang et al. (2024), 'Collective constitutional ai: Aligning a language model with public input'

4.2 Concluding Remarks

In conclusion, this thesis has contributed methods and insights at the interface of AI and social choice theory, demonstrating the potential for LLM-based systems to engage with value pluralism in a more structured and principled manner. By focusing on reflective elicitation of individual preferences and fair, mechanism-based synthesis, this work provides building blocks for developing AI systems and AI-assisted processes that are more sensitive and responsive to the diversity of human values. Continued research in these directions holds promise for improving both the alignment of AI systems and the tools available for collective human decision-making in complex domains.

Bibliography

References are in citation order.

[Bra+16] Felix Brandt et al. *Handbook of computational social choice*. Cambridge University Press, 2016 (cited on page 1).

[Ge+24] Luise Ge et al. 'Axioms for AI Alignment from Human Feedback'. In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 80439–80465 (cited on page 1).

[Con+24] Vincent Conitzer et al. 'Position: social choice should guide AI alignment in dealing with diverse human feedback'. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 9346–9360 (cited on pages 1, 42).

[Fis+24] Sara Fish et al. 'Generative Social Choice'. In: *Proceedings of the 25th ACM Conference on Economics and Computation*. EC '24. New Haven, CT, USA: Association for Computing Machinery, 2024, p. 985. DOI: 10.1145/3670865.3673547 (cited on pages 1, 22–24).

[Sor+24] Taylor Sorensen et al. 'Position: a roadmap to pluralistic alignment'. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 46280–46302 (cited on page 1).

[Noo+18] Ritesh Noothigattu et al. 'A voting-based system for ethical decision making'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018 (cited on page 1).

[BM07] Monika Bütler and Michel André Maréchal. *Framing effects in political decision making: Evidence from a natural voting experiment*. Tech. rep. CESifo Working Paper, 2007 (cited on pages 1, 5).

[Lan20] Hélène Landemore. *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*. Princeton: Princeton University Press, 2020 (cited on page 2).

[FL05] James S Fishkin and Robert C Luskin. 'Experimenting with a democratic ideal: Deliberative polling and public opinion'. In: *Acta politica* 40 (2005), pp. 284–298 (cited on pages 2, 20).

[Fen+23] Shangbin Feng et al. 'From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 11737–11762. DOI: 10.18653/v1/2023.acl-long.656 (cited on page 2).

[RP25] Manon Revel and Théophile Pénigaud. 'AI-Facilitated Collective Judgements'. In: *arXiv preprint arXiv:2503.05830* (2025) (cited on pages 2, 24).

[Sch92] Shalom H Schwartz. 'Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries'. In: *Advances in Experimental Social Psychology/Academic Press* (1992) (cited on page 4).

[Chr+17] Paul F Christiano et al. 'Deep reinforcement learning from human preferences'. In: *Advances in neural information processing systems* 30 (2017) (cited on page 4).

[Ouy+22] Long Ouyang et al. 'Training language models to follow instructions with human feedback'. In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744 (cited on page 4).

[SLH23] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 'Understanding Hidden Context in Preference Learning: Consequences for RLHF'. In: *The Twelfth International Conference on Learning Representations*. 2023 (cited on pages 4–6).

[Cha+24] Souradip Chakraborty et al. 'MaxMin-RLHF: alignment with diverse human preferences'. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 6116–6135 (cited on pages 4, 6).

[Pod+24] Sriyash Poddar et al. 'Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning'. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024 (cited on pages 4–6).

[Fis91] Baruch Fischhoff. 'Value elicitation: Is there anything in there?' In: *American Psychologist* 46.8 (1991), p. 835 (cited on pages 5, 6, 20).

[HDD14] John R Hauser, Songting Dong, and Min Ding. 'Self-reflection and articulated consumer preferences'. In: *Journal of Product Innovation Management* 31.1 (2014), pp. 17–32 (cited on pages 5, 6, 17).

[VVH+20] Niki Ver Donck, Geert Vander Stichele, Isabelle Huys, et al. 'Improving patient preference elicitation by applying concepts from the consumer research field: narrative literature review'. In: *Interactive Journal of Medical Research* 9.1 (2020), e13684 (cited on pages 5, 6).

[HS22] Joseph Hejna and Dorsa Sadigh. 'Few-Shot Preference Learning for Human-in-the-Loop RL'. In: *6th Annual Conference on Robot Learning*. 2022 (cited on page 6).

[Ma+24] Yecheng Jason Ma et al. 'Eureka: Human-Level Reward Design via Coding Large Language Models'. In: *The Twelfth International Conference on Learning Representations*. 2024 (cited on page 6).

[Xie+24] Tianbao Xie et al. 'Text2Reward: Reward Shaping with Language Models for Reinforcement Learning'. In: *The Twelfth International Conference on Learning Representations*. 2024 (cited on page 6).

[Ver+24] Shresth Verma et al. 'Balancing Act: Prioritization Strategies for LLM-Designed Restless Bandit Rewards'. In: *Workshop on Socially Responsible Language Modelling Research*. 2024 (cited on page 6).

[Beh+24] Nikhil Behari et al. 'A Decision-Language Model (DLM) for Dynamic Restless Multi-Armed Bandit Tasks in Public Health'. In: *arXiv preprint arXiv:2402.14807* (2024) (cited on page 6).

[Kwo+23] Minae Kwon et al. 'Reward Design with Language Models'. In: *The Eleventh International Conference on Learning Representations*. 2023 (cited on pages 6, 12).

[FF10] Rowanne Fleck and Geraldine Fitzpatrick. 'Reflecting on reflection: framing a design landscape'. In: *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*. 2010, pp. 216–223 (cited on page 7).

[Koc+18] Rafal Kocielnik et al. 'Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity'. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.2 (July 2018), 70:1–70:26. DOI: 10.1145/3214273. (Visited on 03/12/2024) (cited on page 7).

[Wol+22] Irmtraud Wolfbauer et al. 'A Script for Conversational Reflection Guidance: A Field Study on Developing Reflection Competence With Apprentices'. en. In: *IEEE Transactions on Learning Technologies* 15.5 (Oct. 2022), pp. 554–566. DOI: 10.1109/TLT.2022.3207226. (Visited on 03/12/2024) (cited on page 7).

[AY24] Riku Arakawa and Hiromu Yakura. 'Coaching copilot: blended form of an LLM-powered chatbot and a human coach to effectively support self-reflection for leadership growth'. In: *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 2024, pp. 1–14 (cited on page 7).

[LS06] Sarah Lichtenstein and Paul Slovic. 'The construction of preference: An overview'. In: *The construction of preference* 1 (2006), pp. 1–40 (cited on page 7).

[Mur21] Lala Muradova. 'Seeing the other side? Perspective-taking and reflective political judgements in interpersonal deliberation'. In: *Political Studies* 69.3 (2021), pp. 644–664 (cited on pages 8, 17).

[Awa+18] Edmond Awad et al. 'The moral machine experiment'. In: *Nature* 563.7729 (2018), pp. 59–64 (cited on page 10).

[LK77] J Richard Landis and Gary G Koch. 'The measurement of observer agreement for categorical data'. In: *Biometrics* (1977), pp. 159–174 (cited on pages 13, 15, 17).

[BS99] Patrick D Bridge and Shlomo S Sawilowsky. 'Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research'. In: *Journal of Clinical Epidemiology* 52.3 (1999), pp. 229–235 (cited on page 13).

[Jac12] Paul Jaccard. 'The distribution of the flora in the alpine zone'. In: *New Phytologist* 11.2 (1912), pp. 37–50 (cited on page 14).

[BC06] Virginia Braun and Victoria Clarke. 'Using thematic analysis in psychology'. In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101 (cited on page 14).

[Raf+24] Rafael Rafailov et al. 'From r to Q^*: Your Language Model is Secretly a Q-Function'. In: *First Conference on Language Modeling*. 2024 (cited on pages 22, 23, 25, 34, 35, 41).

[Tes+24] Michael Henry Tessler et al. 'AI can help humans find common ground in democratic deliberation'. In: *Science* 386.6719 (2024), eadq2852 (cited on pages 22, 23, 36, 37, 39, 41).

[CFS17] Vincent Conitzer, Rupert Freeman, and Nisarg Shah. 'Fair public decision making'. In: *Proceedings of the 2017 ACM Conference on Economics and Computation*. 2017, pp. 629–646 (cited on page 24).

[Liu+24] Jiacheng Liu et al. 'Don't throw away your value model! Generating more preferable text with Value-Guided Monte-Carlo Tree Search decoding'. In: *First Conference on Language Modeling*. 2024 (cited on page 24).

[Shi+24] Ruizhe Shi et al. 'Decoding-time language model alignment with multiple objectives'. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 48875–48920 (cited on page 24).

[Cha+25] Souradip Chakraborty et al. 'Collab: Controlled Decoding using Mixture of Agents for LLM Alignment'. In: *The Thirteenth International Conference on Learning Representations*. 2025 (cited on page 24).

[Due+24] Paul Duetting et al. 'Mechanism design for large language models'. In: *Proceedings of the ACM Web Conference 2024*. 2024, pp. 144–155 (cited on page 24).

[Dub+24] Avinava Dubey et al. 'Auctions with llm summaries'. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 713–722 (cited on page 24).

[Eba+24] Soroush Ebadian et al. 'Optimized distortion and proportional fairness in voting'. In: *ACM Transactions on Economics and Computation* 12.1 (2024), pp. 1–39 (cited on pages 27–29, 32).

[Sha71] Lloyd S Shapley. 'Cores of convex games'. In: *International journal of game theory* 1 (1971), pp. 11–26 (cited on page 27).

[ABM19] Haris Aziz, Anna Bogomolnaia, and Hervé Moulin. 'Fair mixing: the case of dichotomous preferences'. In: *Proceedings of the 2019 ACM Conference on Economics and Computation*. 2019, pp. 753–781 (cited on pages 27–29).

[FMS18] Brandon Fain, Kamesh Munagala, and Nisarg Shah. 'Fair allocation of indivisible public goods'. In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. 2018, pp. 575–592 (cited on pages 28, 29).

[Raw71] John Rawls. 'An egalitarian theory of justice'. In: *Philosophical Ethics: An Introduction to Moral Philosophy* (1971), pp. 365–370 (cited on page 28).

[HMS21] Safwan Hossain, Evi Micha, and Nisarg Shah. 'Fair algorithms for multi-agent multi-armed bandits'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 24005–24017 (cited on page 32).

[MJ09] James H Martin and Daniel Jurafsky. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Vol. 23. Pearson-/Prentice Hall Upper Saddle River, 2009 (cited on page 34).

[Xia+23] Shitao Xiao et al. *C-Pack: Packaged Resources To Advance General Chinese Embedding*. 2023 (cited on page 36).

[Gra+24] Aaron Grattafiori et al. 'The llama 3 herd of models'. In: *arXiv preprint arXiv:2407.21783* (2024) (cited on page 36).

[Sad+17] Dorsa Sadigh et al. *Active preference-based learning of reward functions*. 2017 (cited on page 41).

[And+24] Chinmaya Andukuri et al. 'STaR-GATE: Teaching Language Models to Ask Clarifying Questions'. In: *First Conference on Language Modeling*. 2024 (cited on page 41).

[Ste+03] Marco R Steenbergen et al. 'Measuring political deliberation: A discourse quality index'. In: *Comparative European Politics* 1 (2003), pp. 21–48 (cited on page 41).

[RW91] Stuart Russell and Eric Wefald. 'Principles of metareasoning'. In: *Artificial intelligence* 49.1-3 (1991), pp. 361–395 (cited on page 42).

[ZH20] Simon Zhuang and Dylan Hadfield-Menell. 'Consequences of misaligned AI'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15763–15773 (cited on page 42).

[Ge+25] Luise Ge et al. 'Learning Policy Committees for Effective Personalization in MDPs with Diverse Tasks'. In: *arXiv preprint arXiv:2503.01885* (2025) (cited on page 42).

[IB24] Dmitry Ivanov and Omer Ben-Porat. 'Personalized reinforcement learning with a budget of policies'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 11. 2024, pp. 12735–12743 (cited on page 42).

[Bai+22] Yuntao Bai et al. 'Constitutional ai: Harmlessness from ai feedback'. In: *arXiv preprint arXiv:2212.08073* (2022) (cited on page 42).

[Hua+24] Saffron Huang et al. 'Collective constitutional ai: Aligning a language model with public input'. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024, pp. 1395–1417 (cited on page 42).

Appendices

A

Appendix to Chapter 2

A.1 Prompts and ASCII Encodings

A.1.1 Study 1: Multi-Agent Apple Farming Environment

Encodings

We coded each entity in the environment with an ASCII character. The main agent the user was giving feedback on was represented by an "`M`", apples, which the agent could collect for reward, by "`A`", other agents by "`B`", garbage by "`G`", empty tiles by "`.`", and orchard boundaries by "`-`" and "`|`". We also annotated each step in the trajectory with the step number and important actions that occurred. Figure 2.2 shows an example of two timesteps encoded in this way. Notably, all of the information comes from the underlying state of the environment and is programmatically generated.

Prompts

```
For each trajectory, consider the actions of the main agent (M), the
     presence of garbage (G) and apples (A),
and the behavior of background agents (B). Also note that each agent owns
     the quadrant (orchard) that they start in.
Quadrants are defined by the "|" and "-" lines in the ASCII representation
      of the environment.\n

The agent can pick up apples for rewards and clean garbage without reward.
      The agent owns the quadrant (orchard) it starts in but can move to
     other quadrants.
Agents cannot directly interact with one another.

The main agent is colored blue and the other background agents are grey.

You will be presented with 4 scenarios and the user's choices, along with
     their explanations. Your task is twofold:

1. Hypothesize: Based on the user's choices and explanations, identify
     specific decision-making principles they seem to be applying. Look
     for patterns in their reasoning across multiple scenarios and
     formulate a concrete hypothesis about their decision making rules."

2. Offer alternatives: Suggest 2-3 other specific decision-making
     principles the user could apply in these scenarios. These should be
     concrete, actionable, and distinct from the user's apparent approach.

Important guidelines:
- Direct your response to the user, using "you" and "your" when referring
     to their choices and reasoning.
- Keep your responses focused, short, and specific.
- Avoid making moral judgments about the user's choices.

Here are the 4 scenarios and the user's responses:

<trajectories_with_explanations>

Based on these responses, provide your analysis of the user's specific
     decision-making principles and suggest alternative, concrete
     approaches they could consider.
```

Figure A.1: Prompt used to generate the hypothesis and alternative perspectives in Study 1 of Chapter 2.

```
For each trajectory, consider the actions of the main agent (M), the
     presence of garbage (G) and apples (A),
and the behavior of background agents (B). Also note that each agent owns
     the quadrant (orchard) that they start in.
Quadrants are defined by the "|" and "-" lines in the ASCII representation
      of the environment.\n
Each cell is three characters wide and one character tall. A cell with
     only " . " in it is empty.
If a cell has multiple letters, such as "MAG", then the main agent is in
     that cell and there is an apple and garbage in the cell.

The agent can pick up apples for rewards and clean garbage without reward.
      The agent owns the quadrant (orchard) it starts in but can move to
     other quadrants.
Agents cannot directly interact with one another.

All of the information provided by the user is in regard to the main agent
     's (M) actions in the environment.
You will then be asked to label a new trajectory on behalf of the user.

<trajectories_with_explanations>

IMPORTANT: The following hypothesis and user feedback provide crucial
     insights into the labeling criteria. Please pay careful attention to
     these elements as they will significantly influence your labeling
     decision.

Original hypothesis for the user's labels:

<llm_hypothesis>

User feedback on the hypothesis:

<user_feedback>

Please reflect on how this hypothesis and feedback might influence your
     labeling decision.

Now you will be shown a new trajectory and asked to label it as '
     respectful' or 'disrespectful' based on the actions of the main agent
      (M).

Trajectory to label:

<ASCII_encoding_of_trajectory>

Please format your response as follows:
Chain of thought reasoning:

Label:

What follows the "Label:" in your response should be either "respectful"
     or "disrespectful" AND NOTHING ELSE.
```

Figure A.2: Prompt used to generate the trajectory label (reward) in Study 1 of Chapter 2.

A.1.2 Study 2: The Moral Machine

Encodings

Numerical The Moral Machine dataset is transformed into a 26-dimensional vector representation for each scenario by selecting relevant features from the original dataset described in detail here: https://osf.io/wt6mc. These features include scenario structure elements (Intervention, PedPed, Barrier, CrossingSignal), numeric characteristics (NumberOfCharacters, DiffNumberOFCharacters), and 20 character types (e.g., Man, Woman, Pregnant, etc.). Each moral dilemma compares two outcomes (stay vs. swerve), represented by two rows in the dataset. The vectorization process involves creating comparison pairs based on the 'Intervention' column, then calculating the difference between feature values of the two outcomes. This differencing captures the relative changes in features between outcomes, resulting in a 26-dimensional difference vector. Before use in the Multi-Layer Perceptron classifier, these vectors are standardized using StandardScaler to ensure consistent feature scaling.

ASCII The process of creating natural language descriptions for the Moral Machine scenarios involves converting the raw data into a verbal description. Each scenario is described by presenting the basic dilemma of a self-driving car with failed brakes, followed by the two possible outcomes (continuing straight or swerving). The description includes details about the number and types of characters involved in each outcome, their actions (such as crossing legally or illegally), and any relevant attributes (like profession or age). An example is shown in Figure A.3.

```
The brakes of a self-driving car have
failed. The self-driving car can
continue driving straight ahead
or swerve. If the car continues
straight ahead, it will run over
a group of 5 pedestrians who are
crossing against a red do not walk
signal. The group of pedestrians
include:
    - 4 girls
    - A female doctor
If the car swerves, it will run over a
group of 5 pedestrians who are crossing
with a green walk signal. The group of
pedestrians include:
    - 4 boys
    - A male doctor
```

Figure A.3: ASCII encoding of a Moral Machine scenario from Study 2.

Prompts

```
You are an AI assistant analyzing responses to ethical dilemmas based on
     the Moral Machine experiment. In this experiment, participants face
     scenarios where a self-driving car must choose between two
     unavoidable outcomes, each resulting in fatalities. The chosen action
      will definitely cause the deaths described.

You will be presented with 6 scenarios and the user's choices, along with
     their explanations. Your task is twofold:

1. Hypothesize: Based on the user's choices and explanations, identify
     specific decision-making principles they seem to be applying. Look
     for patterns in their reasoning across multiple scenarios and
     formulate a concrete hypothesis about their decision making rules."

2. Offer alternatives: Suggest 2-3 other specific decision-making
     principles the user could apply in these scenarios. These should be
     concrete, actionable, and distinct from the user's apparent approach.

Important guidelines:
- Direct your response to the user, using "you" and "your" when referring
     to their choices and reasoning.
- Keep your responses focused, short, and specific.
- The scenarios always involve unavoidable fatalities.
- Users are forced to choose between two outcomes; there is no option to
     avoid harm entirely.
- Avoid making moral judgments about the user's choices.
- If the user says "go straight", it means the same as "continue on the
     current path", "keep left", or other similar instructions. The option
      on the left side of the screen is always the default "stay" path.
- If the user says "The vehicle should swerve" or "go to the right", "
     change lanes", or other similar instructions, it means the same thing
     . The swerve option is always shown on the right side of the screen.

Here are the 6 scenarios and the user's responses:

<trajectories_with_explanations>

Based on these responses, provide your analysis of the user's specific
     decision-making principles and suggest alternative, concrete
     approaches they could consider.
```

Figure A.4: Prompt used to generate the hypothesis and alternative perspectives in Study 2 of Chapter 2.

```
You are an AI tasked with predicting how a specific user would label
     ethical dilemmas involving autonomous vehicles. Your goal is to
     accurately mimic the user's decision-making process based on their
     previous choices.

Here are the user's previous decisions:

<trajectories_with_explanations>

Based on these decisions, here's a hypothesis about the user's decision-
     making process:

<llm_hypothesis>

The user provided this feedback on the hypothesis:

<user_feedback>

IMPORTANT: This hypothesis and feedback are crucial for understanding the
     user's decision-making process. They provide insights into the user's
      ethical framework and reasoning that may not be immediately apparent
      from their choices alone. Use this information to guide your
     prediction, but remember to critically evaluate it against the user's
      actual choices.

Now, predict how this user would label the following new scenario:

<ASCII_encoding_of_trajectory>

Your response should be in this format:
Chain of thought reasoning: [Step by step reasoning]

Prediction: [Your predicted label: ONLY either the word "stay" or "swerve"
      AND NOTHING ELSE]
```

Figure A.5: Prompt used to generate the trajectory label (reward) in Study 2 of Chapter 2.

A.2 Descriptions of Participant Decision Making Features

A.2.1 Study 1: Multi-Agent Apple Farming Environment

The 12 features we identified are:

1. **Stays in Own Quadrant:** Whether the main agent stays in the orchard (quadrant) they own.
2. **Interferes With Others:** This feature evaluates a relatively wide range of behaviours, including if the agent appeared to attempt to block another agent, following another agent, or getting too close to another agent.
3. **Task Completion:** Whether the agent works toward or completes the task the participant thought the agent should be doing. For example, an agent picking up all apples or garbage in their quadrant.
4. **Picks Up Own Garbage:** Whether the agent picks up garbage in their own orchard (quadrant)
5. **Picks Up Others' Garbage:** Whether the agent picks up garbage in other agents orchards

6. **Tit for Tat Behaviour:** Some participants explained that a given behaviour was respectful if another agent had done it to them first, such as entering their quadrant.
7. **Taking Others' Apples**: Whether the agent picks up apples in orchards it does not own.
8. **Eats Own Apples:** Whether the agent eats any apples in their own quadrant.
9. **Picks Up Garbage Before Apple:** Whether the agent picks up garbage before eating apples in their own or other quadrants.
10. **Efficiency:** Whether the agent moves around without collecting apples or garbage or makes repetitive, futile movements
11. **Time in Others' Quadrants:** The duration of time spent in other agents' orchards.
12. **In Quadrant While Owner was Gone:** If the agent entered another agent's quadrant while they were gone.

A.2.2 Study 2: The Moral Machine

The 9 features we identified are:

1. **Human vs. Animal Priority:** Whether human lives are prioritized over animal lives.
2. **Traffic Rule Compliance:** The consideration of the legality of pedestrians' actions.
3. **Minimizing Casualties:** The aim to minimize the total number of fatalities.
4. **Age Prioritization:** The preference for saving younger people over older people.
5. **Survival Probability:** The consideration of the likelihood of survival for different individuals based on factors like physical fitness.
6. **Maintaining Original Path:** The preference for the vehicle to stay on its original course rather than swerving.
7. **Societal Value:** The consideration of perceived societal value or potential future contributions of individuals.
8. **Passenger Safety Priority:** The prioritization of the safety of the vehicle's passengers over pedestrians or other road users.
9. **Gender:** The consideration of the gender of potential victims in the decision-making process.

B

Appendix to Chapter 3

B.1 Prompts

```
You are generating a consensus statement that represents the views of
     multiple participants.
Your task is to continue the statement in a way that addresses the issue
     and considers all participants' opinions. Be concise and keep the
     statement short (less than 50 tokens) and focused. ONLY WRITE THE
     STATEMENT AND NOTHING ELSE.

Issue:
<issue>

Participants' opinions:
<opinion_1>
.
.
.
<opinion_n>

Consensus statement:
```

Figure B.1: Reference policy prompt used in Chapter 3.

```
You are generating a statement that represents the views of a single
     participant.
Your task is to continue the statement in a way that addresses the issue
     and considers ONLY this participant's opinion. Be concise and keep
     the statement short (less than 50 tokens) and focused. ONLY WRITE THE
      STATEMENT AND NOTHING ELSE.

Issue:
<Issue>

Participant's opinion:
<opinion>

Statement reflecting ONLY this participant's opinion:
```

Figure B.2: Agent policy prompt used in Chapter 3.

```
You are helping to fix ONLY the ending of a generated statement.

VERY IMPORTANT: If the statement ending is already complete and well-
     formed, DO NOT modify it at all.

Your task is to:
1. DO NOT change any part of the statement except the last few sentences
     if they have issues
2. Look for and fix ONLY these issues at the end of the statement:
   - Remove repetition in the final sentences
   - Complete any unfinished final sentence that can be completed easily
   - Remove any incomplete final sentence that cannot be meaningfully
     finished
3. Keep the changes minimal and focused only on the ending
4. DO NOT add any new information or opinions
5. DO NOT modify anything except problematic sentences at the end
6. If the statement is already complete and well-formed, return it EXACTLY
      as provided

Here is the statement:

<statement>

Please examine ONLY the ending of this statement. If the ending is already
      well-formed and complete, return the statement unchanged.

If the ending has issues (repetition, incomplete sentences), fix ONLY
     those issues at the end while keeping the rest of the statement
     exactly the same.

Return ONLY the statement without explanations or comments.
```

Figure B.3: Brush up prompt used in Chapter 3.

B.2 Credit Assignment Results for Gemma

Reference policy prompt	User policy prompt	Sequence	Z-Score
User food profile: empty	User food profile: vegetarian	I am having chicken/tofu en chila das tonight . Then I am going to meet up with some friends .	3.69
User location profile: empty	User location profile: lives in a cold climate	I ' m going to the beach/mountains this weekend to surf/ski . I need to buy some new clothes .	1.01, 2.86
User time profile: empty	User time profile: morning	I am about to eat some food . I am going to have spaghetti/pancakes . I will use my phone to order it .	3.13
User opinion: empty	User opinion: Favors stricter gun control laws.	Implementing background checks that are less/more strict for gun purchases is essential . Also , my favorite color is orange .	1.77

Table B.1: Credit assignment results for `Gemma 2 9b Instruction-Tuned`. Darker green indicates larger Z-score. Z-score column is for altered tokens. Alterations are represented by "<misaligned>/<aligned>".

B.3 Scenarios and Sample Consensus Statements

Table B.2: Summary of Agent Opinions and Method Consensus on Scenario 1: "*Are we living in a golden age of democracy?*"

Source	Opinion / Consensus Statement
	Individual Agent Opinions
Agent 1	No, I think the golden age of democracy is long gone. I think a system where the first past the post wins is not working and we need to move to a model of proportional representation which would give more people the feeling that their voices were being heard. On the subject of voices, I'm strongly of the opinion that we have become a society where the loud voices of minority groups are able to impose their views on the rest of the population which to me is no democracy at all.
Agent 2	Worldwide democracy is more present than it's ever been in history. So yes, compared to previous ages in history I believe we are. Although that's not to say we can't improve - many countries still don't operate democratically, and in the ones that do, corruption is rife.
Agent 3	Yes, we are living in a golden age of democracy as democracy is of the people.
Agent 4	Not at all. The notion of democracy is being used for personal gains of those in government, and the system is manipulated. Around the world there is a considerable amount of oppression and lack of democractic values.
Agent 5	Compared to some parts of the world such as Russia and China which are actively regressing and reverting back to archaic ways of controlling their people, most Western countries are living through comparatively decent times, although problems still exist.
	Consensus Statements from Methods
Best of N	While acknowledging the imperfect nature of democracies worldwide, and recognising instances of regression and manipulation, we contend that, compared to historical periods and several non-democratic countries, the majority of Western countries and a sizeable portion of the global population currently enjoys a relatively high level of democratic freedoms; however, the overarching consensus leans towards a need for more inclusive and accountable forms of governance.
Finite Lookahead	Given the varying perspectives, we acknowledge that the definition of democracy can be subjective and dependent on geographical and cultural contexts, yet still, we recognize the existence of significant flaws in the current democratic systems worldwide, including instances of corruption, oppression, and manipulation, leading to the erosion of trust in governments and the democratic process. While some countries have made significant progress in consolidating democratic institutions, others have regressed, highlighting the need for continuous improvement and reform to ensure the preservation and strengthening of democratic values and principles. Therefore, we conclude that the current state of democracy worldwide is complex and multifaceted, with areas of improvement necessary to fully realize the ideals of democracy for all citizens, and acknowledge the presence of democratic systems in many parts of the world, particularly in the Western hemisphere, which, despite their imperfections, remain the best available option for promoting the common good and protecting individual rights and freedoms.
Prompted Habermas Machine	We acknowledge that the concept of a golden age of democracy is complex and contentious. While democracy has made significant progress in many parts of the world, we also recognize that fundamental challenges persist, including issues with representation, corruption, and the suppression of marginalized voices. These challenges highlight the need for a more critical examination of our democratic systems and a commitment to ongoing reform and improvement.
Habermas Machine	Although in the past we've had high hopes of a better world after the horrors of WWII and the fall of the Iron Curtain, democracy is in danger worldwide and may never reach its full potential. The Western world has poor democratic values, and even though democracy is spreading worldwide it is being overshadowed by the loud voices of minority groups.
Beam Search	We are not living in a golden age of democracy, but rather in an age where democracy is being challenged from within and from outside, and it is up to all of us to ensure that the principles of democracy are upheld.

Table B.3: Summary of Agent Opinions and Method Consensus on Scenario 2: "*Should the UK's ties to Europe be stronger or weaker?*"

Source	Opinion / Consensus Statement
	Individual Agent Opinions
Agent 1	When we was in Europe we had good trade with them , The decision to leave was very bad for united kindom . We need to put the vote again to the British public i am sure this time the decision would be to remain
Agent 2	The natural evolution of our species has been to grow into ever bigger "tribes". Families ruled by their patriarchs became tribes ruled by elders became countries ruled by governments. It made sense that countries would evolve separately since they were geographically separate with no means of communication. Now our world is so connected, it is inevitable that we evolve into ever larger units such as the United States and the European Union. Eventually we will become a multi-planetary species ruled by an Earth government. To sever ties with Europe is a step in the wrong direction.
Agent 3	Although we did exit EU few years ago, we are geographically and economically in proximity to most EU countries. Several geographic, financial, political and economical parameters are intertwined with our neighbouring countries and, it would be advantageous to be in good relations to fully harness our economic, political, and financial facilities.
Agent 4	I believe the UK's ties to Europe should be stronger, as it would make trade deals easier to negotiate. This would allow us to benefit from a larger array of goods, which would make our imports cheaper. I also believe that closer ties with Europe in terms of immigration policies should occur as our immigration policy doesn't coincide with other nations around us. I believe a more united Europe would help all countries grow more through the movement of free labour, goods and services and more.
Agent 5	I think they should be stronger, as all the countries of Europe, except us, are part of the EU, and it makes Britain more isolated. Countries are stronger I believe when they work together with other countries and are open to the influence of other countries. I believe that leaving the EU was a step backwards and done for the wrong reasons. We are not the country we were 100 or even 50 years ago and the world is changing so much that we cannot afford to not be part of Europe.
	Consensus Statements from Methods
Best of N	Given the participants' opinions, we conclude that in order to achieve economic prosperity and streamline international relations, the UK's ties to Europe should be strengthened through collaborative efforts in trade, immigration, and economic policies while embracing the benefits of European integration, mutual understanding, and cooperation.
Finite Lookahead	The UK's ties to Europe should be strengthened to promote economic growth, facilitate trade, and foster greater unity among European nations. This can be achieved through mutually beneficial agreements and policies that reflect the shared interests of the UK and its European neighbours, leading to a more harmonious and prosperous region for all. Let's work together to build a stronger, more integrated Europe for the benefit of all its citizens, including the British people. Let's put the past behind us and look forward to a brighter, more collaborative future for the UK and Europe. Let's recognize the benefits of unity and work towards creating a more cohesive and successful Europe for generations to come.
Prompted Habermas Machine	We propose strengthening ties with Europe to foster greater economic co-operation, promote a more unified Europe, and address global challenges. However, we acknowledge that the path to this goal should include a reevaluation of the decision to leave the EU, allowing for a more informed and inclusive discussion about our relationship with Europe. Moreover, we recognize the long-term implications of European integration as part of a broader process of global unification, which may ultimately lead to a more interconnected and cooperative world.
Habermas Machine	The UK's ties to Europe should be stronger. This is because, although the UK did leave the EU, we are geographically and economically in proximity to most EU countries. Several geographic, financial, political and economical parameters are intertwined with our neighbouring countries and, it would be advantageous to be in good relations to fully harness our economic, political, and financial facilities.
Beam Search	The UK's ties to Europe should be stronger, and it would be beneficial for the country to be a part of the European Union or a similar entity that fosters cooperation and unity among its member states. This would allow for more efficient trade and economic growth, as well as greater collaboration on issues such as immigration, security, and environmental policy.

Table B.4: Summary of Agent Opinions and Method Consensus on Scenario 3: "*Is it okay to use animals for food?*"

Source	Opinion / Consensus Statement
	Individual Agent Opinions
Agent 1	Yes, as human beings are natural meat eaters. However, it is morally wrong to keep animals in cruel conditions and use them for food. On the whole, farming is an unethical practise and many farming methods should be banned.
Agent 2	I think that it is ok to use animals for food because I enjoy eating meat. I don't think it is ok if the animal is endangered and I do think that there should be certain factors in place to ensure that they are well kept.
Agent 3	In my opinion it is ok to use animals for food. They provide nutrition and vitamins to use and I think it is just the basic food chain. I do understand why people choose to be vegatarian or vegan, but I do not agree with them trying to force their views upon others.
Agent 4	Humans have used animals for food since we first evolved as a species. Animals provide a great deal of nutrition that the human body needs to be healthy. It is certainly true that we now have more alternative foods available so that it can be possible to live healthily without consuming meat - though this does usually rely on taking nutritional supplements. I think that we should continue to diversify food production to provide people with meat alternatives, but also we should continue to improve the welfare of animals that are eaten for food and try to do this sustainably to minimise the impact this has on the animals wellbeing and on the land / environment too.
	Consensus Statements from Methods
Best of N	While considering the ethical and nutritional aspects of using animals for food, we agree that it is acceptable to consume meat in moderation, provided that animals are raised and slaughtered humanely, and that sustainable and environmentally-friendly farming methods are implemented to minimize harm.
Finite Lookahead	The use of animals for food is acceptable, however it should be done in a way that respects the welfare of the animals and considers the impact on the environment, with the aim of achieving sustainable food production methods that benefit both human and animal well-being, and also providing alternatives for those who choose not to consume meat. This includes improving farming practices, reducing the amount of waste and pollution, and promoting education on responsible food production and consumption practices. All sides of the issue should be taken into account, and the goal should be to achieve a balance between food security, animal welfare, and environmental sustainability.
Prompted Habermas Machine	We acknowledge that humans have traditionally used animals for food, and we recognize the importance of improving animal welfare and minimizing environmental impact to ensure sustainable practices. We understand that some consumers prefer eating meat, and we encourage diversifying food production options to include alternatives that can meet the nutritional needs of humans. Importantly, we emphasize the need to prohibit inhumane farming practices and actively work towards improving animal conditions, which aligns with our collective commitment to animal well-being and environmental stewardship.
Habermas Machine	Humans have used animals for food since we first evolved as a species. Animals provide a great deal of nutrition that the human body needs to be healthy. It is certainly true that we now have more alternative foods available so that it can be possible to live healthily without consuming meat - though this does usually rely on taking nutritional supplements. We should continue to diversify food production to provide people with meat alternatives, but also we should continue to improve the welfare of animals that are eaten for food and try to do this sustainably to minimise the impact this has on the animals wellbeing and on the land / environment too.
Beam Search	It is generally acceptable to use animals for food, as long as they are treated with respect and care, but we must also consider the impact of our food choices on the environment and animal welfare.