Jiseon Kim
PhD Candidate at KAIST

AI for Pluralistic Societies: From Individual Value Judgments to Institutional Value Dynamics

As AI systems increasingly participate in decisions that hinge on human values, a central challenge emerges: *whose* values they reflect and *how* these values shift across contexts, personas, and institutional settings. My research addresses this challenge by examining how diverse values appear and interact across social life—from individual moral judgments to the institutional processes that shape policy. **I investigate how AI can understand and align with these pluralistic values, advancing value-aware reasoning and supporting more context-sensitive responses.**

To pursue this agenda, I study how value differences arise in both everyday moral reasoning and the political domains where collective decisions are made. At the micro level, I analyze how large language models respond to changes in framing or persona to probe their capacity to represent diverse perspectives. At the macro level, I examine how values are embedded in complex domains such as legislation, where group interests interact through evolving texts and social networks. This work spans **(1) foundational model training**, **(2) value-focused evaluation**, **(3) legislative analysis**, and **(4) frameworks for pluralistic value modeling**, aiming to develop AI methodologies that are **transparent, socially grounded, and responsive to diverse communities**.

1. **Training:** Training language models to learn **rich semantic representations**;
2. **Evaluation:** Benchmarking AI understanding of **diverse social values**;
3. **Application:** Analyzing **diverse interests** and **values shaping legislative processes**; and
4. **Pluralistic Value Modeling:** Designing frameworks for **pluralistic alignment** and deliberation.

1. Learning Rich Semantic Representations

Human expression spans a wide, continuous range of subtle variations that is not fully captured by discrete categorical labels. Yet the data needed to model such fine-grained variation is often limited. To address this, my research moves beyond discrete supervision by **(1) mapping categorical labels into continuous, fine-grained value dimensions** and **(2) crafting synthetic textual data augmentation for nuanced variation**.

Continuous Emotional Representation Learning [1]. Standard categorical emotion models often miss the gradients that underlie social values. In this work, I contributed to developing a learning framework that embeds categorical labels into a continuous space, enabling models to capture fine-grained variations such as emotional intensity and nuance rather than coarse discrete states. This shift provides a basis for aligning AI with the fluid and dynamic structure of human sentiment.

Nuanced Synthetic Textual Data Augmentation [2]. To better distinguish intrinsic meaning across diverse expressions, I contributed to developing a contrastive learning framework that integrates data augmentation with curriculum learning. By progressively introducing augmented data from easier to harder levels, the model learns to focus on core semantic meaning rather than superficial patterns. This approach yields more sample-efficient and generalizable representations.

2. Benchmarking Understanding of Diverse Social Values

While training models on rich data is essential, it is equally critical to rigorously evaluate their alignment with the pluralistic values of people. My research identifies a significant alignment gap: whereas human values are inherently complex and context-dependent, models often default to narrow, specific perspectives that fail to reflect the diversity of the people they serve. I address this by developing robust benchmarks to assess **(1)cultural bias**, **(2) demographic-specific moral judgments**, and **(3) multi-perspective social reasoning**.

Quantifying Cultural Bias [3, 4]. Social biases differ across cultures, but existing benchmarks remain predominantly Western. To address this gap, I developed an adaptable evaluation framework that can be applied to diverse cultural contexts. Demonstrating this approach, I constructed Korean cultural-bias datasets to capture local stereotypes often missed by general benchmarks. This framework provides methodology to assess whether models generalize fairness across diverse cultural groups.

Persona-Sensitive Moral Judgments [5]. Moral values are often shared broadly, yet the prioritization of these values varies across demographics. Using moral dilemma task (e.g., the trolley problem), I investigated whether models can distinguish and replicate these demographic nuances. My findings reveal that current LLMs struggle to maintain consistent demographic personas, often exhibiting inconsistent behavior across contexts. This limitation underscores the difficulty models face in representing the distinct preferences of specific social groups.

Jiseon Kim
PhD Candidate at KAIST

Multi-Perspective Social Reasoning [6, 7]. Social intelligence relies on the ability to interpret a single situation through diverse perspectives. I evaluate this capability by probing whether models can accurately adapt to shifting viewpoints while maintaining coherent social reasoning. Drawing on cognitive science tasks like Theory of Mind, I identify limitations in how models track distinct belief states as observational perspectives change [6]. In moral dilemmas, I further analyze how judgments shift when queries probe different evaluative dimensions—ranging from moral rightness to expected social norms and practical actions [7]. These findings reveal that model reasoning is highly sensitive to framing, raising critical questions about their robustness and coherence across different social layers.

3. Analyzing Diverse Interests and Values Shaping Legislative Processes

Beyond evaluating values within models, I apply AI to analyze how diverse values interact in the real world. I specifically focus on legislative systems, as they serve as the primary institutional arenas where competing societal preferences and pluralistic values are negotiated and translated into law. However, the scale and complexity of these processes often obscure whose values are effectively represented. To bring transparency to these dynamics, my work focuses on **(1) bill similarity**, to trace how legal provisions evolve across policy domains, and **(2) interest-group preferences**, to identify whose values shape legislative outcomes.

Bill Similarity [8]. Legislation consists of structured documents encoding diverse stakeholder interests, which undergo complex editing interactions with one another, such as merging and splitting. I develop data-driven methods to measure bill-level similarity by capturing these structural and domain-specific semantic features. By constructing networks based on these similarities, I provide a fundamental framework to trace the flow of legislation. This allows for a structural analysis of how specific legal provisions evolve and connect across different policy domains.

Interest-group Preference [9]. Various stakeholders actively participate in lawmaking to advocate for their interests. To understand this landscape, I build large-scale datasets that capture the specific positions—supporting, opposing, amending, or monitoring—that groups take on legislation. By analyzing these datasets, I track the preferences and values of diverse groups and examine their lobbying strategies, such as cross-industry coalitions and conflicts. This quantitative approach empirically uncovers whose interests are reflected in the final legislative outcomes.

4. Future Directions: Pluralistic Value Modeling

Building on my research agenda—from representation learning and value evaluation to sociopolitical analysis—my long-term goal is to develop AI systems that can recognize, adapt to, and help negotiate pluralistic human values. As LLMs increasingly participate in socially consequential tasks, future work must move beyond output-level alignment toward models that are transparent, socially grounded, and capable of reasoning about heterogeneous communities. To pursue this vision, I focus on three directions: **(1) mechanistic interpretability of social reasoning, (2) pluralistic alignment**, and **(3) multi-agent value negotiation.**

Interpretability of Social Reasoning. Existing work on social or moral reasoning primarily evaluates model outputs, offering limited insight into the internal mechanisms that produce them. I will develop methods to reveal how LLMs encode normative concepts—such as fairness or political ideology—within their latent representations and attention pathways. By mapping these internal structures, I can move from black-box behavioral alignment to mechanistic understanding, enabling more targeted and reliable interventions.

Pluralistic Alignment. Traditional alignment approaches assume a unified behavioral standard, neglecting the diversity of human values. My work addresses this by modeling divergences between human and model preferences. For example, my prior work on factor-level preferences [10] shows how disentangling fine-grained value dimensions improves alignment. Building on this, I will develop value-conditioned models that adapt to community norms or deliberative goals, and that can explicate the assumptions underlying their recommendations.

Multi-Agent Value Negotiation. Real-world decision making is inherently multi-perspectival, from everyday moral debates to legislative negotiations. To model these dynamics, I will develop multi-agent LLM systems where each agent is parameterized by a distinct value profile or stakeholder interest. By examining how agents justify positions, challenge assumptions, and revise proposals, I can identify core patterns of agreement, conflict, and compromise. This approach enables empirical analysis of how value dimensions shape deliberation, including which arguments drive outcomes or how amendments shift burdens across groups.

Jiseon Kim
PhD Candidate at KAIST

References

[1] Sungjoon Park, **Jiseon Kim**, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, Alice Oh, "**Dimensional emotion detection from categorical emotion**" *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[2] Seonghyeon Ye, **Jiseon Kim**, Alice Oh, "**Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning**" *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[3] {**Jiseon Kim**, Jiho Jin, Nayeon Lee, Hanual Yoo}*, Alice Oh, Hwaran Lee, "**KoBBQ: Korean Bias Benchmark for Question Answering**" *Transactions of the Association for Computational Linguistics (TACL)*, 2024.

[4] HyperCLOVA X Team (**Jiseon Kim** as Contributor), "**HyperCLOVA X Technical Report**", *Technical Report*, 2024.

[5] {**Jiseon Kim**, Jea Kwon, Luiz Felipe Vecchietti}*, Alice Oh, Meeyoung Cha, "**Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment**", *Bidirectional Human-AI Alignment Workshop @ ICLR*, 2025.

[6] Chani Jung, Dongkwan Kim, Jiho Jin, **Jiseon Kim**, Yeon Seonwoo, Yejin Choi, Alice Oh, Hyunwoo Kim, "**Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Model**", *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[7] **Jiseon Kim**, Jea Kwon, Luiz Felipe Vecchietti, Wenchao Dong, Alice Oh, Meeyoung Cha, "**Mind the Gap: LLM Actions vs. Human Social Understanding in Moral Dilemmas**", *Women in Machine Learning Workshop @ NeurIPS*, 2025.

[8] **Jiseon Kim**, Elden Griggs, In Song Kim, Alice Oh, "**Learning Bill Similarity with Annotated and Augmented Corpora of Bills**" *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[9] **Jiseon Kim**, Dongkwan Kim, Joohye Jung, In Song Kim, Alice Oh, "**Measuring Interest Group Positions on Legislation: An AI-Driven Analysis of Lobbying Reports**", *Women in Machine Learning Workshop @ NeurIPS 2024 (Preparing Submissions to Science)*, 2025.

[10] {Juhyun Oh, Eunsu Kim}*, **Jiseon Kim**, Wenda Xu, Inha Cha, William Yang Wang, Alice Oh, "**Uncovering Factor Level Preferences to Improve Human-Model Alignment**", *Conference on Empirical Methods in Natural Language Processing (EMNLP-Findings)*, 2025.