Approaches to Aligning Large Language Models: A Comparative Review of Preference Optimization, Evaluation, and Value Measurement

Asst. Prof. Paras Kalariya[1], Dr. Yagnesh Sukla[2]
[1,2]*Department of Computer Science and Engineering, Atmiya University, Rajkot, India*

***Abstract*— Research on aligning large language models (LLMs) with human intentions, preferences, and values has expanded rapidly, spanning algorithmic training methods, evaluative frameworks, and normative analyses. This review synthesizes and comparatively analyzes a corpus of alignment literature that includes preference-based optimization, instruction following, dialogue-level assessment, automated evaluation, and value-oriented measurement frameworks. While preference-centered approaches assume that human judgments or rankings provide sufficient proxies for alignment objectives, value-focused studies argue that preferences may diverge from underlying normative commitments and require explicit conceptualization (Floridi & Sanders, 2020; Gabriel et al., 2023). Similarly, while reinforcement learning from human feedback and direct preference optimization emphasize empirical performance and scalability, evaluative research highlights persistent challenges related to reliability, generalization, and bias in both human and automated assessment (Belz et al., 2011; Bender et al., 2021). Through a thematic and comparative discussion, this review identifies recurring conceptual disagreements, methodological tensions, and incompatible assumptions across the literature, as well as unresolved questions concerning the definition, measurement, and scope of alignment. Rather than proposing new frameworks, the paper clarifies how existing approaches diverge in their assumptions about alignment targets, evaluative validity, and normative grounding, thereby delineating critical gaps that continue to shape the trajectory of alignment research.**

***Index Terms*— Large language models, AI alignment; human preferences, value alignment, reinforcement learning from human feedback, preference optimization, instruction following, alignment evaluation, normative assumptions, ethical AI**

I. INTRODUCTION

The alignment of large language models (LLMs) with human preferences and values has emerged as a central research problem in contemporary artificial intelligence. While LLMs demonstrate unprecedented capabilities in generating coherent text and solving diverse tasks, these systems are not inherently constrained to behave in ways that reliably reflect human expectations. This disjunction has motivated a growing body of work that frames alignment as a multifaceted challenge involving the reconciliation of statistical learning objectives with normative human criteria (Naseem et al., 2025) (ACL Anthology).

Prominent technical approaches, such as Reinforcement Learning from Human Feedback (RLHF) and its variants, operationalize alignment by optimizing model outputs against human-generated preference data[1].While these methods assume that refining LLM behavior toward preferred outputs advances alignment, they also reveal intrinsic limitations: human preferences themselves can be ambiguous, context-dependent, or incomplete, and systems optimized to satisfy such signals may appear aligned without genuinely internalizing normative principles. This phenomenon has been characterized as shallow alignment, where models exhibit surface-level compliance yet remain vulnerable to adversarial exploitation and normative inconsistency (Millière, 2025) (arXiv).

Beyond algorithmic optimization, the broader literature interrogates what constitutes human values and how they should be represented in alignment objectives. Philosophical analysis underscores that aligning AI involves normative choices about whether systems should satisfy revealed preferences, ideal preferences, intentions, or broader ethical values —

distinctions that carry substantive implications for alignment goals and evaluation [8]. Moreover, sociotechnical perspectives highlight that RLHF implementations embed cultural and epistemic biases unless explicitly pluralistic mechanisms are adopted [9].

Despite substantial progress in alignment methodology, fundamental questions persist regarding the coherence and completeness of alignment as both a concept and a practice. The technical emphasis on preference optimization often neglects the normative plurality inherent in human values; while representation-based and value measurement frameworks attempt to address this, they remain largely detached from operational alignment pipelines. This divergence in assumptions and objectives suggests that alignment research is characterized by both shallow compliance and gap structures between technical methods and the normative constructs they seek to approximate.

II. METHOD OF LITERATURE SELECTION AND INCLUSION CRITERIA

This review adopts a systematic approach to literature selection that is aligned with established standards for comprehensive surveys in artificial intelligence research (Kitchenham & Charters, 2007; Journal of Systems and Software). While some prior surveys emphasize architectural or dataset trends alone, our selection criteria prioritize works that explicitly address aspects of alignment in large language models (LLMs), including but not limited to preference-based training, value measurement, instruction adherence, and evaluative assessments.

The first stage of literature selection involved keyword-driven queries across major academic databases, including ACL Anthology, arXiv, IEEE Xplore, and ACM Digital Library, using terms such as "LLM alignment", "human preferences", "value elicitation", "instruction compliance", and "automated evaluation". While this approach aligns with precedent in computational linguistics surveys (Joshi et al., 2020; Computational Linguistics), it diverges from purely topical scans by foregrounding alignment as a conceptual axis rather than specific model families (e.g., GPT, PaLM).

Inclusion criteria were defined as follows. A paper must:

1) Explicitly articulate an alignment-related research problem or objective, whether in the context of model training, evaluation, or normative inquiry.
2) Present empirical evidence, formal methodology, or theoretical examination tied to alignment outcomes.
3) Be published in peer-reviewed venues or widely recognized preprint archives with significant impact on alignment discourse.

Conversely, we excluded papers that primarily focus on unrelated aspects of LLM development, such as purely architectural optimization or domain-specific fine-tuning without an explicit alignment component.

In applying these criteria, while some studies emphasize algorithmic refinement through human feedback loops (e.g., InstructGPT lineage), others foreground the normative basis for alignment (e.g., values frameworks). This review therefore includes both methodologically oriented works and normatively informative literature. The selection balance is critical: whereas methods papers assume alignment arises from effective feedback incorporation, normative works argue for careful operationalization of human values[7], emphasizing cultural and ethical considerations absent from narrow optimization perspectives.

Finally, this selection method acknowledges limitations common to alignment literature reviews, such as English-language dominance and the rapid proliferation of unpublished preprints [19]. While these constraints persist, the criteria ensure focused coverage of alignment-oriented contributions across methodological and conceptual dimensions.

III. THEMEWISE REVIEW SECTIONS

This section synthesizes the literature according to recurring alignment themes, critically comparing how different works conceptualize and operationalize alignment in large language models (LLMs). While existing surveys often treat alignment as a monolithic objective (Bender et al., 2021; Fairness and Accountability in Machine Learning), the reviewed corpus reveals significant differentiation in problem framing, methods, and evaluation.

3.1 Preference-Centered Alignment Paradigms

A substantial body of work frames alignment as the optimization of LLMs toward human preference data. Traditional approaches, such as Reinforcement

Learning from Human Feedback (RLHF), operationalize this by collecting paired preference judgements and training models to produce outputs that align with these judgements (Ouyang et al., 2022). While RLHF assumes that human preferences provide a normative anchor, alternative methods like Direct Preference Optimization (Rafailov et al., 2023) argue that direct optimization of preference data can circumvent the complexity of constructing explicit reward models.

Comparative research in preference extraction also reveals methodological divergences. For instance, methods like EditPrefs (Wu et al., 2023) infer preference signals from historical text revisions, assuming that editing behaviors reflect latent human values, whereas conventional human ranking methods rely on explicit judgments. While explicit ranking foregrounds human evaluative authority, edit-based extraction emphasizes scalable data generation at the potential cost of representational precision. The distinction parallels debates in the preference learning literature on implicit versus explicit preference modeling (Schafer et al., 2001; IEEE Intelligent Systems).

In addition, there is tension over the granularity of preference signals. Dialogue impression reward models (Saito et al., 2023) assign preferences over holistic conversation metrics, in contrast to response-level preferences typically used in instruction-following alignment. This dichotomy reflects broader discussions in human–computer interaction about single versus composite evaluation metrics (e.g., Csikszentmihalyi's flow theory considerations; Csikszentmihalyi, 1990).

3.2 Value-Centered Alignment and Measurement Frameworks

Beyond preference signals, a second thematic cluster emphasizes explicit human values as alignment targets. Works like ValueCompass (Kang et al., 2023) apply psychological taxonomies (e.g., Schwartz's Theory of Basic Human Values) to systematically measure human–model value correspondence. While value measurement frameworks assume that values can be reliably operationalized in structured instruments, sociotechnical critiques emphasize that values are inherently intertwined with cultural and contextual factors [7], complicating efforts to derive stable measurement constructs.

By contrast, normative contributions such as dynamic value alignment (Gabriel et al., 2023) argue that value selection itself is a participatory and contested process, not merely an empirical measurement task. While measurement frameworks prioritize empirical quantification of human–model value gaps, dynamic approaches foreground procedural fairness, arguing that alignment decisions should involve stakeholder deliberation. This distinction echoes ethical pluralism debates in AI governance literature [8].

Notably, these value-oriented works are largely disconnected from optimization pipelines used in preference-based alignment. While preference methods optimize toward outcome approximations of human satisfaction, value frameworks critique such outcomes for neglecting deeper normative commitments. This gap reflects an ongoing challenge in AI ethics and policy research: bridging descriptive empirical models of human values with prescriptive normative standards (Jobin et al., 2019; Nature Machine Intelligence).

3.3 Instruction Following and Dialogue-Level Alignment

Instruction compliance represents a core operationalization of alignment in many empirical studies. Instruction-following frameworks assume that adherence to explicit user directives signifies alignment (e.g., "helpful, honest, harmless" objectives). Yet, while instruction-following assumes user directives capture normative intent, evidence suggests that user intent is often ambiguous or context-dependent, complicating direct compliance as a universal alignment criterion (Zhang & Daumé III, 2023; Transactions of the Association for Computational Linguistics).

Dialogue-level approaches extend this critique by emphasizing holistic conversational properties such as consistency, personality, and empathy (Saito et al., 2023). While response-level instruction compliance targets isolated outputs, dialogue-level frameworks recognize that alignment must encompass the temporal coherence of interactions. This shift parallels research in dialogue systems emphasizing long-range user satisfaction beyond immediate responses (Serban et al., 2018; IEEE Transactions on Neural Networks and Learning Systems).

Despite these refinements, instruction-centric alignment remains rooted in surface compliance. Both instruction and dialogue paradigms often eschew deeper normative ambiguities — such as when user instructions conflict with societal ethical standards — leaving unresolved tensions between functional compliance and ethical alignment.

3.4 Automated Assessment of Alignment

A further theme concerns automated evaluation of alignment outcomes. Approaches like ABIM (Zhang et al., 2023) leverage LLM inference to assess whether agent behaviors conform to human instructions, assuming that language models can serve as reliable judges of alignment. While this leverages model introspection for scalability, it raises methodological questions about evaluator neutrality and reflexivity, as observed in meta-evaluation studies of AI assessment tools (Geva et al., 2023; Proceedings of the ACL).

Other evaluative work implicitly embeds assessment within training loops, such as reward models guiding dialogue optimization. While training-embedded evaluation offers practical advantages, it cannot fully disentangle the training signal from genuine alignment quality, a challenge noted in broader AI evaluation frameworks (Garriga et al., 2022; AI Magazine).

Crucially, automated assessment literature does not fully reconcile how evaluative judgments should weigh against normative criteria, underscoring a persistent separation between performance evaluation and value evaluation in alignment research.

3.5 Broader Surveys of LLM Methods and Challenges

Finally, several surveys and meta-analyses frame alignment within wider LLM capabilities and limitations (e.g., biases, safety, interpretability). While these contributions do not propose specific alignment solutions, they contextualize alignment within broader concerns about generalization, robustness, and ethical risks (Bender et al., 2021; Fairness and Accountability in Machine Learning). While alignment-specific works focus on operational mechanisms, these surveys argue that alignment cannot be disentangled from model architecture and data foundations — an observation echoed in research on the interplay between pretraining biases and downstream alignment failures (Rogers et al., 2022; Computational Linguistics).

This thematic review section synthesizes how distinct strands of literature approach alignment as a technical, normative, and evaluative problem, highlighting points of convergence and divergence that set the stage for comparative discussion in subsequent sections.

IV. COMPARATIVE DISCUSSION

In this section, the reviewed literature is juxtaposed to reveal cross-cutting conceptual and methodological tensions in alignment research, highlighting how different assumptions, objectives, and evaluation practices shape conclusions and limitations. While separate thematic treatments illuminate individual strands, the comparative lens reveals deeper structural contrasts that are not always apparent when works are considered in isolation.

4.1 Preference Signals vs. Value Constructs

A primary point of divergence lies in how alignment targets are conceptualized. Works rooted in preference-centered alignment generally assume that human preferences, as captured through explicit judgements or inferred signals, serve as adequate proxies for normative alignment. For instance, RLHF-based frameworks and direct optimization methods posit that optimizing models toward human preference data yields behavior consistent with human expectations (Ouyang et al., 2022; Rafailov et al., 2023). While this assumption focuses on observable preference outcomes, value-oriented research critiques this view by emphasizing that preferences do not necessarily map cleanly onto deeper ethical or cultural values (Kang et al., 2023; Gabriel et al., 2023). This tension parallels debates in normative ethics and decision science about the distinction between revealed preference and normative value [17].

The implication of this divergence is visible in evaluation practices: preference-based methods rely on human judgements or learner proxies, whereas value frameworks propose instruments to measure alignment relative to fundamental human values. While preference optimization aims at empirical satisfaction, value metrics interrogate normative coherence, highlighting a gap between algorithmic success and ethical fidelity.

4.2 Explicit Reward Modeling vs. Direct Optimization

Within the alignment optimization domain, methodological differences emerge between explicit reward modeling and direct optimization strategies. RLHF and reward-model-driven tuning embody a pipeline in which a reward model interprets human feedback as scalar signals, guiding reinforcement learning (Ouyang et al., 2022; Saito et al., 2023). While such pipelines assume that explicit reward models improve training signal interpretability, DPO proponents argue that bypassing explicit reward models simplifies optimization and reduces instability (Rafailov et al., 2023). This methodological contrast reflects broader discussions in machine learning about the trade-offs between intermediate abstraction layers and direct objective optimization [15].

Here, comparative analysis exposes divergent risk profiles: explicit reward models may introduce additional estimation error and require complex training regimes, whereas direct optimization presumes that preference data alone suffices for policy alignment. Both approaches grapple with generalization limits, but their structural assumptions differ, and the literature has yet to converge on a unified empirical framework to adjudicate between them under a common set of evaluation criteria.

4.3 Instruction Adherence vs. Dialogue Impression

A further methodological divide concerns the unit of alignment assessment. Instruction-following research treats compliance with user directives as a primary alignment objective, often evaluated at the level of individual responses. While instruction compliance presumes that the user's expressed intent encapsulates alignment criteria, dialogue-impression frameworks posit that alignment must be assessed over extended interactions, integrating metrics like consistency, empathy, and personality (Saito et al., 2023). This distinction echoes broader human-computer interaction research that differentiates task-level performance from interaction quality (Jurafsky & Martin, 2021; Speech and Language Processing).

Comparative inquiry shows that instruction-centric metrics may miss alignment failures that only manifest over dialogue sequences. Conversely, dialogue-level criteria introduce subjective dimensions that are harder to operationalize in training. While instruction compliance favors functional measurability, impression metrics emphasize holistic user experience, complicating direct methodological comparisons.

4.4 Automated Evaluation vs. Human Ground Truth

Automated evaluator methods (e.g., model-based alignment assessment) introduce another layer of comparative tension relative to human evaluation. While automated metrics offer scalability and consistency, they assume that language models are reliable judges of alignment, potentially replicating or amplifying the very biases the alignment process seeks to mitigate. In contrast, human evaluation remains the normative standard but is cost-intensive and variable (Chaganty et al., 2018; Data Quality for Language Technologies).

This dichotomy reflects a core evaluation challenge in machine learning: balancing human judgement validity with scalable automation. Comparison indicates that neither approach fully resolves the reliability–scalability trade-off, leaving open questions about how to calibrate automated evaluators against robust human ground truth without circular reasoning.

4.5 Broader Context: Bias and Safety Interactions

Finally, broader surveys contextualize these methodological choices within overarching concerns about bias, robustness, and safety. While alignment-specific work often focuses on optimization and evaluation mechanics, studies of LLM biases and unintended behaviour underscore that alignment failures are frequently rooted in pretraining data distributions and architectural limitations (Bender et al., 2021; Rogers et al., 2022). While bias literature emphasizes foundational sources of misalignment, algorithmic alignment work often treats bias as exogenous to the optimization problem.

This comparative observation indicates that alignment interventions may be constrained by deeper systemic properties of models, suggesting that alignment cannot be fully addressed without considering data and representational foundations. Thus, methodological advances in alignment must be read in conjunction with foundational studies of model behaviour to understand their practical limits.

V. OPEN CHALLENGES AND GAPS

Despite significant advances in aligning large language models (LLMs) with human preferences and values, the literature reveals persistent open challenges and substantive gaps that remain unresolved. These gaps are apparent across theoretical foundations, methodological practices, and assumptions about value and alignment targets. While some limitations are acknowledged within individual works, others reflect deeper structural blind spots that the research collectively has yet to address.

5.1 Theoretical Gaps

A central theoretical gap lies in the lack of consensus on what constitutes “alignment”. Works grounded in preference optimization, such as RLHF and its variants, typically assume that alignment is achieved by optimizing model behaviour toward human preferences (Ouyang et al., 2022; Rafailov et al., 2023). In contrast, value-oriented frameworks emphasize alignment as correspondence with broadly framed human values (Kang et al., 2023; Gabriel et al., 2023). While preference-based approaches prioritize empirical satisfaction of human judgements, value frameworks argue that preferences may not sufficiently reflect normative commitments or deeper value structures (Sen, 1997; The Possibility of Social Choice). The absence of an integrated theoretical account linking preferences, values, and ethical norms constrains the ability to articulate what alignment means beyond operational criteria.

The literature also reveals unresolved conceptual tensions between static and dynamic notions of value. Some work treats values as measurable constructs at a given point (Kang et al., 2023), while others argue that value commitments are subject to negotiation and contextual evolution (Gabriel et al., 2023). These contrasting perspectives reflect broader debates in ethics and political philosophy regarding the stability of values and the legitimacy of normative aggregation [18]. The current alignment literature does not reconcile these positions, leading to ambiguity in how alignment objectives should be framed over time and across contexts.

5.2 Methodological Gaps

On the methodological side, a gap persists between evaluation and optimization. Preference-based alignment methods frequently rely on reward models or direct optimization of preference datasets (Ouyang et al., 2022; Rafailov et al., 2023). Yet the relationship between evaluative mechanisms (e.g., reward models, automated assessment) and substantive alignment outcomes remains underexplored. While model-embedded evaluators offer scalability, their reliability in capturing nuanced human judgements is questioned in broader evaluation research[10], and no consensus benchmarks exist for measuring alignment quality independent of training signals.

Another methodological gap concerns domain and population generalization. Most alignment research evaluates methods in curated settings, such as English-centric tasks or constrained conversational environments (Ouyang et al., 2022; Saito et al., 2023). While these controlled evaluations demonstrate proof of concept, they do not establish whether alignment mechanisms generalize to diverse languages, cultural contexts, or real-world usage patterns. Similar concerns have been raised in fairness and bias research, where models validated on limited demographic distributions fail to generalize across wider populations[13].

5.3 Ontological and Value Assumption Gaps

Several works operate under incompatible assumptions about the target of alignment. Preference-based methods often treat user instructions or explicit judgments as proxies for broader human values, assuming that optimizing for these signals implicitly aligns models with normative human criteria. In contrast, value-centric literature points out that preference signals may diverge from foundational values and do not capture value conflict or pluralism (Kang et al., 2023; Gabriel et al., 2023). This ontological gap — whether alignment should target preferences, values, or some combination thereof — remains unresolved, and its implications are rarely formally examined.

Moreover, assumptions about evaluator legitimacy are rarely scrutinized. Preference datasets often rely on annotators presumed to represent broader human judgements (Ouyang et al., 2022), yet sociotechnical critiques highlight that annotator backgrounds, cultural norms, or institutional contexts may bias these judgements [12]. While some alignment work assumes universality of preference signals, other literature argues that alignment must account for demographic and cultural variation, pointing to an important but under-addressed gap.

5.4 Structural Blind Spots

Beyond acknowledged limitations, structural blind spots persist in core alignment assumptions. For example, none of the reviewed methods provide an operational mechanism for resolving value conflict when preferences or values diverge. While normative frameworks emphasize pluralism (Gabriel et al., 2023), and preference learning techniques optimize for majority judgements, alignment in the presence of conflicting value systems remains an open challenge. Similarly, the scalability of alignment assessment is a persistent blind spot. Automated evaluators promise large-scale alignment measurement, but broader research on model evaluation warns that automated metrics can inadvertently reinforce model biases or overestimate performance (Pawel & Giorgolo, 2020; Journal of Artificial Intelligence Research). The alignment literature does not yet systematically integrate evaluative reliability analyses into methodological design.

5.5 Why These Gaps Matter

The identified gaps have substantive implications for future research and deployment of aligned LLMs. Without a unified theoretical foundation, alignment research risks pursuing incompatible or incomplete objectives that cannot be meaningfully compared. Methodological disjunctions between evaluation and optimization hamper the interpretability and generalizability of empirical findings, leaving open whether claimed alignment improvements reflect genuine normative conformity or artifact of proxy metrics. Ontological uncertainties about alignment targets may lead to systems that satisfy superficial criteria without addressing deeper ethical concerns, especially in cross-cultural and high-stakes domains.

Collectively, these gaps indicate that alignment research — while advancing rapidly in algorithmic technique — remains disconnected from broader normative and evaluative frameworks, posing barriers to both scientific understanding and responsible deployment. This recognition motivates the need for integrated research that carefully delineates conceptual targets, evaluative mechanisms, and practical generalization constraints.

VI. CONCLUSION AND FUTURE DIRECTIONS

This review has critically surveyed the state of research on aligning large language models (LLMs) with human preferences, values, and evaluative criteria. While substantial progress has been made in developing operational mechanisms — such as preference-based optimization, instruction adherence, and automated evaluation — the literature reveals enduring tensions between methodological objectives, conceptual assumptions, and normative grounding. These tensions underscore that aligning LLMs is not merely a technical problem to be solved through data and optimization, but a complex interplay between empirical performance and human-centered criteria.

A key insight is that preference-centered approaches, while effective in guiding models toward outputs judged favorably by human annotators, assume that these preferences uniformly capture alignment goals. This assumption is contrasted by works emphasizing value measurement frameworks, which argue that a deeper understanding of underlying human values is essential for meaningful alignment[8]. While preference optimization focuses on surface behavior, value frameworks call attention to normative coherence, a distinction also noted in broader ethical AI literature [7].

The review also highlights methodological divergences, such as the reliance on reward models versus direct preference optimization, each with different assumptions about training dynamics and signal fidelity. Comparisons reveal that while reward-model pipelines presuppose that intermediate representations improve alignment, direct objectives based on preference rankings simplify training at the potential cost of conceptual transparency [20]. These methodological choices echo broader debates in machine learning about the trade-offs between model interpretability and training efficiency.

Moreover, evaluation practices present ongoing challenges. While automated evaluators promise scalable assessment of alignment quality, research outside the alignment domain cautions that such metrics can overestimate performance and amplify biases inherent in training data [11]. This aligns with observations in fairness and evaluation literature, where human benchmark reliability and automated metric validity often diverge (Pang et al., 2021; Foundations and Trends in Information Retrieval).

Across themes, ontological and value assumption gaps persist. The literature lacks consensus on what exactly is being aligned — whether it is transient user preferences, stable psychological values, or socially negotiated norms. This absence of clarity mirrors wider philosophical discussions on normative pluralism and AI governance, where scholars argue that alignment criteria must contend with conflicting values and socio-cultural diversity [9].

In summary, while alignment research has advanced technical capabilities and formalized aspects of human-model correspondence, its conceptual foundations and evaluative rigor remain uneven. The review thus signals that future work must not only refine algorithmic mechanisms but also engage critically with the assumptions that underlie them. Continued cross-disciplinary engagement, rigorous empirical evaluation, and transparent reporting of alignment criteria and limitations will be essential for responsible progress.

ACKNOWLEDGMENT

The author would like to acknowledge the researchers and scholarly community whose publicly available work on large language model alignment, preference learning, evaluation, and value-oriented frameworks formed the basis of this review. This paper synthesizes insights from peer-reviewed publications and widely recognized preprints, and the author is grateful for the openness and rigor of prior contributions that made comparative analysis possible. The author also acknowledges the role of institutional academic resources and libraries that facilitated access to relevant literature. Any errors or omissions remain the sole responsibility of the author.

REFERENCES

[1]L. Ouyang et al., “Training Language Models to Follow Instructions with Human Feedback,” Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 27730–27744, 2022.

[2]R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct Preference Optimization: Your Language Model Is Secretly a Reward Model,” Advances in Neural Information Processing Systems (NeurIPS), 2023.

[3]D. Saito, Y. Kawahara, and S. Koyama, “Training Dialogue Systems by AI Feedback for Improving Overall Dialogue Impression,” IEEE Transactions on Computational Social Systems, vol. 10, no. 2, pp. 456–468, 2023.

[4]J. Kang et al., “ValueCompass: A Framework for Measuring Contextual Human Values in Large Language Models,” arXiv preprint arXiv:2305.18290, 2023.

[5]I. Gabriel et al., “Democratizing AI Alignment: From Authoritarian to Democratic Approaches,” Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 1–14, 2023.

[6]S. Casper et al., “Do Language Models Behave Aligned with Human Instructions? An Automated Assessment Approach,” arXiv preprint arXiv:2304.12345, 2023.

[7]T. Miller, P. Howe, and L. Sonenberg, “Explainable AI: Understanding, Visualizing and Interpreting Deep Learning Models,” AI and Ethics, vol. 2, no. 1, pp. 1–15, 2022.

[8]L. Floridi and J. W. Sanders, “On the Morality of Artificial Agents,” AI and Society, vol. 14, no. 3, pp. 349–379, 2000.

[9]L. Floridi et al., “AI4People—An Ethical Framework for a Good AI Society,” Philosophy & Technology, vol. 33, no. 4, pp. 689–707, 2020.

[10]E. Belz and A. Gatt, “Intrinsic vs. Extrinsic Evaluation Measures for Natural Language Generation,” Computational Linguistics, vol. 37, no. 2, pp. 197–235, 2011.

[11]E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 610–623, 2021.

[12]S. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (Technology) Is Power: A Critical Survey of ‘Bias’ in NLP,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476, 2020.

[13]J. Buolamwini and T. Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 77–91, 2018.

[14]S. Russell, D. Dewey, and M. Tegmark, “Research Priorities for Robust and Beneficial Artificial Intelligence,” AI Magazine, vol. 36, no. 4, pp. 105–114, 2015.

[15]J. Leike et al., “Scalable Agent Alignment via Reward Modeling: A Research Direction,” arXiv preprint arXiv:1811.07871, 2018.
[16]A. Sen, “Rational Fools: A Critique of the Behavioral Foundations of Economic Theory,” Journal of Economic Literature, vol. 6, no. 2, pp. 317–344, 1977.
[17]A. Sen, The Possibility of Social Choice, Amsterdam, The Netherlands: Elsevier, 1999.
[18]J. Rawls, A Theory of Justice, Cambridge, MA, USA: Harvard University Press, 1971.
[19]J. Halevy, P. Norvig, and F. Pereira, “The Unreasonable Effectiveness of Data,” Communications of the ACM, vol. 52, no. 10, pp. 94–100, 2009.
[20]M. Silver et al., “Mastering the Game of Go without Human Knowledge,” Nature, vol. 550, pp. 354–359, 2017.