JOHANNES KEPLER
UNIVERSITY LINZ

Author
Giovanni Filomeno
12345678

Submission
Seminar in AI (Master)
(365.205)

March 26, 2025

Your Language Modell is Secretly a Reward Model

Seminar Report

Abstract This seminar report reviews techniques for aligning Large Language Models (LLMs) with human preferences, contrasting the established Reinforcement Learning from Human Feedback (RLHF) pipeline with the simpler, direct approach of Direct Preference Optimization (DPO). It examines the theoretical underpinnings, benefits, and challenges of these methods, including data coverage and the complexity of human preferences, while highlighting recent developments in the field.

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Contents

Abstract **3**

1. Story Summary **3**

2. Introduction **6**

3. Background Pipeline **6**
3.1 Supervised Fine-Tuning (SFT) 6
3.2 Human-Preference Collection and Reward Modeling 7
3.3 Reinforcement Learning from Human Feedback (RLHF) 7
3.4 Direct Preference Optimization (DPO) 7
3.5 Other Preference Optimization Approaches 8

4. Literature Review: Evolving Preference Alignment Techniques **8**
4.1 Alternate Preference Parameterizations: Beyond Scalar Rewards 8
4.2 Self-Play and Iterative Fine-Tuning for Data Acquisition 9
4.3 Handling Data Limitations: Partial Coverage and Non-Transitivity 9

5. Conclusion **10**

References **11**

Abstract

Recent advances in large-scale language models have highlighted the challenges of aligning these models with human intent and preference. Among numerous alignment strategies, Reinforcement Learning from Human Feedback (RLHF) has attracted particular attention due to its demonstrated success in steering model behavior. However, RLHF pipelines often involve multiple components (e.g., a reward model and an RL algorithm like PPO) that can be costly, unstable, or prone to reward hacking. Direct Preference Optimization (DPO) offers a simpler alternative by implicitly learning a reward function and extracting an optimal policy in closed form, which enables more stable and computationally lightweight training. In this literature review, the motivation, theoretical underpinnings, and empirical performance of DPO relative to other preference-based approaches are examined. Also, the implications of DPO for safer and more controllable text generation are discussed, and open questions regarding scalability, partial observability, and robustness to non-transitive human preferences are highlighted. Throughout, additional perspectives from contrastive learning, reward modeling, and iterative feedback loops are incorporated, synthesizing insights that inform both research and practical deployment of preference-aligned language models.

1. Story Summary

1. **What is the central question?**
   If the language models can be effectively aligned with human preferences by focusing on methods like Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO). Key methodological choices that ensure stable training, computational efficiency, and robust, human-aligned outputs, while mitigating issues like reward hacking and instability, are the focus of this inquiry.

2. **Why is this question important?**
   As highlighted in the provided papers, large language models, when trained solely on next-token prediction, can exhibit unintended and potentially harmful behaviors. Aligning these models with human preferences is crucial for safety, ethical considerations, and ensuring their utility in real-world applications. The ability to steer these models towards desired behaviors is essential for responsible deployment. As stated in [**lambert2022illustrating**], "what makes a 'good' text is inherently hard to define as it is subjective and context dependent." Therefore, it is important to align the models with human preferences.

3. **What evidence/data (variables) are needed to answer this question?**
   To answer this question, diverse preference datasets are essential. These datasets should consist of human annotations (or proxies) that compare model outputs. Key variables include:

   - Preference accuracy (win rates) against baselines.
   - Quantitative measures of reward model calibration.
   - Qualitative measures of alignment (e.g., toxicity reduction, helpfulness)
   - Data related to training stability (e.g., convergence rates, variance).
   - Computational efficiency (training time and resource usage).
   - Scalability metrics.

- Metrics on robustness to distribution shift.
- Metrics on the model's ability to handle intransitive human preferences.

4. **What methods are used to get this evidence/data?**
   From literature, the following methods have been employed:
   - Human annotation: Gathering preference labels from human annotators by presenting them with pairs of model outputs.
   - Automated evaluation: Using models (e.g., GPT-4) as proxies for human judgment.
   - RL Pipelines: Implementing RLHF, which involves training a reward model and using it to optimize the language model's policy.
   - DPO: Directly optimizing the policy using preference data through a classification loss, as described in [8].
   - Constrastive Preference Learning Learning directly from preferences without explicitly learning a reward function, as shown in [14].
   - Self-Play Preference Optimization: Using the model to generate its own training data.
5. **What analyses must be applied for the data to answer the central question?**
   The following analyses are crucial to answer the central question:
   - Statistical evaluations: Measuring preference accuracy, reward model loss curves, and convergence rates.
   - User studies: Gathering qualitative feedback on model outputs.
   - Comparing analyses: Investigating the impact of hyperparameters, reference policies, and reward designs on performance. Additionally, the comparison can be made on model level, comparing the performance of RLHF, DPO, and other preference-based methods.
   - Analysis of computational cost.
6. **What evidence/data (values for the variables) were obtained?**
   Empirical studies have yielded evidence of improved preference satisfaction scores and reduced toxic content in RLHF- and DPO-tuned models. Additionally, higher user satisfaction has been observed across various tasks, including summarization, dialogue, and instruction following. However, performance variability exists, depending on dataset size and preference consistency. Furthermore, [8] shows that DPO is more computationally efficient than RLHF. The study in [14] shows that contrastive preference learning can be a viable alternative to RLHF.
7. **What were the results of the analyses?**
   Analyses have demonstrated that RLHF reliably improves alignment, though it is susceptible to reward hacking and high computational overhead. Conversely, DPO has shown comparable or improved performance with a more stable and computationally lighter training pipeline, with its success contingent upon accurate implicit reward parameterization. Alternative approaches, such as contrastive preference learning, have demonstrated promise in bypassing explicit reward modeling, and Self-Play Preference Optimization shows that the model can be used to generate it's own training data.

8. **How did the analyses answer the central question?**
   The analyses have provided evidence that methods combining preference data with sound optimization strategies can effectively align models with human intent. Notably, DPO offers a closed-form alignment solution, suggesting that complex reinforcement learning may not always be necessary for achieving robust preference alignment.

9. **What does this answer tell us about the broader field?**
   These findings indicate that preference-based fine-tuning is essential for controlling large language models. Moreover, simpler and more stable alignment techniques, such as DPO, represent viable alternatives to complex reinforcement learning approaches. Consequently, future research efforts should prioritize efficient and safe AI deployment, with a focus on addressing scalability, distribution shift, and the handling of inconsistent human preferences. The field is progressing towards more diverse and efficient methods for aligning LLMs with human preferences.

10. **Did the paper answer the question satisfactorily? Why (not)?**
    YES, within the constraints of current empirical data and theoretical frameworks, the reviewed documents suggest that RLHF and DPO, along with other methods, have made substantial progress in aligning LLMs. However, ongoing research is necessary to address challenges related to scalability, mitigate distribution shift issues, handle potential intransitive or inconsistent human preferences, further explore the theoretical foundations of preference-based learning, and investigate alternative methodologies.

2. Introduction

Large language models (LLMs) have rapidly evolved to achieve near-human or even superhuman performance on tasks such as reading comprehension, summarization, and instruction following. Yet, the question of *how* to align these models with human interests, safety requirements, and behavioral norms remains at the forefront of AI research. Purely maximizing next-token likelihood on large text corpora often fails to produce outputs that reflect the specific preferences or values of intended end-users. Reinforcement Learning from Human Feedback (RLHF) [4] has thus emerged as one influential approach, coupling human-annotated comparisons with a reward model to iteratively steer policy gradients toward human-aligned outputs. However, RLHF pipelines commonly rely on multiple training phases (reward model learning, policy optimization) and can suffer from high computational overhead or reward gaming [6].

An alternative family of methods focuses on learning directly from preference data without explicit reinforcement loops. Notably, *Direct Preference Optimization (DPO)* [8] proposes to reparameterize the reward function in a way that yields an optimal policy in closed form, effectively turning the traditional RLHF pipeline into a more stable, single-stage procedure. Recent studies show that DPO can match or exceed the performance of RLHF on tasks such as summarization or sentiment control, while mitigating some of the complexities inherent in policy-gradient methods like PPO [9]. Meanwhile, numerous variants of preference-based alignment continue to emerge, ranging from contrastive preference learning [14] to frameworks that address non-transitive human judgments [12].

This literature review examines the conceptual motivations behind both RLHF and direct preference approaches, highlighting their benefits and trade-offs. Section 2 outlines the background pipeline for RLHF, including typical steps of supervised fine-tuning and reward model training. Section 3 delves deeper into the DPO algorithm, illustrating how its *closed-form* solution bypasses the need for iterative on-policy sampling. In Section 4, we survey other preference-based strategies and contextualize them within the broader field of human-aligned language modeling. Finally, the Conclusion discusses outstanding questions related to scaling, model interpretability, and handling ambiguities or inconsistencies in human-labeled feedback. Through this lens, we underscore that the real challenge lies not only in *whether* LLMs can be preference-tuned, but in choosing *how* best to combine computational tractability with fidelity to diverse and often complex human values.

3. Background Pipeline

Aligning large language models (LLMs) with human intentions often begins with a multi-step *preference optimization pipeline*. In practice, this pipeline can follow a trajectory from *supervised fine-tuning* (SFT) on high-quality demonstrations to more involved *reinforcement* or *direct optimization* steps. Below we summarize the stages typically found in the Reinforcement Learning from Human Feedback (RLHF) approach and then briefly contrast it with the Direct Preference Optimization (DPO) paradigm.

3.1 Supervised Fine-Tuning (SFT)

During SFT, a pre-trained language model (such as GPT-like or encoder-decoder transformers) is tuned on a collection of high-quality input–output pairs. Often, these pairs

come from human-written examples or carefully curated data that reflects the desired style, task performance, or safety profile. The objective is the standard cross-entropy loss over the ground-truth tokens, effectively shifting the model from a general-purpose distribution to one more aligned with an intended domain or task. A notable advantage of SFT is that it is straightforward to implement, provided enough high-quality demonstrations exist. However, it may fail to capture nuanced user preferences or subtle policy constraints beyond the scope of the explicit training data. This SFT model often serves as the starting point or **reference policy** (π_{ref}) for subsequent preference tuning stages.

3.2 Human-Preference Collection and Reward Modeling

Once a model is supervised fine-tuned, practitioners typically collect a *human preference dataset* ($\mathsf{D} = \{(\mathsf{x}, \mathsf{y}_w, \mathsf{y}_l)\}$, where y_w is preferred over y_l for prompt x) by asking annotators to compare pairs of model-generated outputs for the same context or prompt. For instance, given a request such as "*Summarize the following paragraph,*", the model may produce two different completions, and annotators select which is more satisfactory, e.g., in terms of correctness or clarity. These relative judgments form a dataset of preferences, which can then be used to train a *reward model* (RM). The RM learns a scalar score $\mathsf{r}_\phi(\mathsf{x}, \mathsf{y})$ for each candidate output so that higher scores correlate with "more preferred" completions, often modeled using frameworks like the Bradley–Terry model [2]. Unlike purely supervised data, this pairwise preference data can be more efficient for complex tasks, because it is often easier for humans to rank outputs than to produce perfect references [10].

3.3 Reinforcement Learning from Human Feedback (RLHF)

In the RLHF setting, the policy, typically initialized from the SFT model (π_{ref}), is further optimized using the trained reward model r_ϕ as the reward signal within a reinforcement learning framework. Notably, policy-gradient methods like Proximal Policy Optimization (PPO) [9] are commonly employed. These methods often involve training both the policy itself (the actor) and an associated **value model** (the critic) to estimate expected future rewards and stabilize updates. The overall objective maximizes the expected reward $\mathsf{E}_{\mathsf{y}\sim\pi_\theta(\mathsf{y}|\mathsf{x})}[\mathsf{r}_\phi(\mathsf{x}, \mathsf{y})]$ while incorporating a KL-divergence penalty $-\beta \mathsf{D}_{\mathsf{KL}}(\pi_\theta(\cdot|\mathsf{x})\|\pi_{\text{ref}}(\cdot|\mathsf{x}))$ to prevent the optimized policy π_θ from diverging too drastically from the initial reference policy π_{ref}. Over multiple rounds of *on-policy* sampling and updates, the model aims to produce outputs that score higher under the learned reward function, thus aligning it more closely with user preferences. While effective in practice, RLHF can exhibit instability due to the challenges of on-policy sampling at scale, potential reward gaming, or calibration issues with the reward model [**bib:gao2023scaling**].

3.4 Direct Preference Optimization (DPO)

Direct Preference Optimization (DPO) proposes a *closed-form* approach for transforming pairwise preferences directly into policy updates [**bib:rafailov2023direct**]. Rather than explicitly construct a reward function and then apply RL, DPO leverages an analytical relationship between the optimal RLHF policy and the underlying reward function. It shows that under the typical KL-constrained RLHF objective, the optimal policy can be derived directly from the preference data using a simple classification-like loss, effectively bypassing the need for explicit reward modeling and iterative on-policy rollouts.

The model can be updated directly using the offline preference dataset D. Early results suggest that DPO simplifies the training pipeline significantly while maintaining or exceeding the performance of standard RLHF in tasks such as summarization, dialogue, and instruction following. However, DPO's theoretical derivation relies on assumptions such as the Bradley–Terry preference model, potentially facing challenges when human preferences are inconsistent or non-transitive [12].

3.5 Other Preference Optimization Approaches

Beyond RLHF and DPO, various methods aim to address preference alignment. Contrastive Preference Learning [**bib:zhao2023contrastive**], for example, moves away from explicit reward modeling, instead leveraging pairwise ranks directly in a contrastive objective function applied to the policy. Others explore value-based frameworks or adopt multi-stage processes that iteratively refine outputs with human feedback. While these strategies often share the common theme of "learning from partial order data," their implementation details vary widely.

In the following sections, we delve deeper into Direct Preference Optimization and alternative preference-based algorithms. We compare their theoretical foundations, typical outcomes, and potential pitfalls, setting the stage for a broader understanding of how human feedback can be effectively harnessed to guide large language models.

4. Literature Review: Evolving Preference Alignment Techniques

While the pipeline described in Section 3 represents a standard framework for aligning LLMs with preferences, recent literature has introduced several refinements and alternatives that either extend or diverge from Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO). This section examines three interconnected themes shaping current research: (1) alternate parameterizations of preference models moving beyond simple scalar rewards, (2) the use of self-play and iterative schemes to dynamically acquire preference data, and (3) methods designed to handle the complexities of real-world feedback, such as partial data coverage and non-transitive preferences.

4.1 Alternate Preference Parameterizations: Beyond Scalar Rewards

A significant limitation of standard reward modeling, including the implicit reward in DPO [**bib:rafailov2023direct**] derived from the Bradley–Terry model [2], is its reliance on a single scalar reward function which may struggle to capture the full complexity or potential inconsistency of human preferences. Several works aim to overcome this:

- **Identity Preference Optimization (IPO)** [1]: Instead of mapping preferences to an intermediate reward function, IPO proposes to optimize the policy to directly match empirical preference probabilities. By learning the mapping $(x, y_w, y_l) \to P(y_w \succ y_l|x)$ without assuming an underlying scalar reward, IPO aims to directly maximize the likelihood of observed human choices. This approach might inherently

handle intransitive preferences better than scalar-based methods and potentially mitigates the reward shifting issue related to the partition function normalization implicit in DPO.

- **Kahneman–Tversky Optimization (KTO)** [5]: Drawing inspiration from behavioral economics, KTO reformulates the preference objective using principles from prospect theory. It models human utility not just based on absolute outcomes but also relative gains and losses compared to a reference point, aiming to capture known psychological biases. While early results show promise in reproducing human-like decision patterns, KTO introduces additional complexity into the optimization landscape and may require careful calibration.

These approaches signify a trend towards richer, potentially more robust representations of human preferences compared to the traditional single-scalar reward paradigm common in RLHF.

4.2 Self-Play and Iterative Fine-Tuning for Data Acquisition

Contrasting with methods relying solely on a static, offline preference dataset (D), another research direction focuses on dynamic data generation through iterative refinement or self-play mechanisms. The core idea is that as the LLM policy (π_θ) improves, the preference data used for alignment should ideally adapt to reflect its new capabilities and failure modes.

- **Self-Play Preference Optimization (SPO)** [11]: This framework formalizes alignment as a two-player game where the policy aims to maximize its win rate against opponent policies (often previous versions of itself or a reference policy). In each iteration, responses are sampled from the current policy, evaluated (e.g., using an updated preference model or comparing against reference outputs), and this feedback is used to further refine the policy. SPO treats preference maximization as the direct objective, potentially avoiding intermediate reward modeling.
- **Iterative RLHF/DPO Variants**: Although less formalized than SPO, practical implementations sometimes involve periodically refreshing the preference dataset by collecting new human judgments on outputs from the updated policy, or using AI feedback [7].

Such iterative schemes offer the potential benefit of improved data coverage over the evolving policy's output distribution and continuous adaptation. However, they introduce significant engineering challenges (e.g., maintaining diverse sampling, efficient feedback collection) and risk exacerbating *distributional shift* if the policy explores regions poorly understood by the preference model or human annotators. This closely relates to the exploration-exploitation trade-off inherent in reinforcement learning.

4.3 Handling Data Limitations: Partial Coverage and Non-Transitivity

Real-world human feedback often deviates from idealized assumptions. Two key challenges are non-transitivity and partial data coverage:

- **Non-Transitivity and Contextuality**: As noted by Tversky [12], human preferences can be intransitive (A $>$ B, B $>$ C, but C $>$ A) or highly context-dependent. Standard models like Bradley–Terry inherently assume transitivity. To address this,

recent works explore context-aware reward functions or multi-dimensional utility representations [13], allowing preferences to shift based on the specific prompt or scenario. While potentially more realistic, these models increase complexity.

- **Partial Coverage and Robustness**: Offline datasets (D) inevitably provide only partial coverage of the vast space of possible model outputs. Optimizing solely on this limited data can lead to overfitting or poor generalization, especially if the policy learns to exploit artifacts in the preference data or reward model. To mitigate this, researchers propose incorporating *pessimism* or *conservative* updates [3]. These techniques penalize deviations from the reference policy or add regularization based on data density, aiming to prevent the model from confidently selecting actions in poorly covered regions of the state-action space, thus ensuring safer extrapolation. This connects directly to the coverage issues highlighted as potential failure modes for offline methods like DPO when assumptions break down.

Discussion and Open Challenges. Overall, the literature reflects two primary thrusts: (1) simplifying the alignment pipeline by moving away from explicit RL (e.g., DPO, IPO, CPL [**bib:zhao2023contrastive**]), and (2) enriching the representation and acquisition of preference data to better capture the nuances of human values (e.g., KTO, SPO, context-aware models). Significant open questions remain at the intersection of these trends. Is it feasible to develop a unified framework that combines the stability of direct methods with the adaptive data generation of iterative schemes? How can we best manage the fundamental trade-off between alignment *stability* (reducing risks like reward hacking) and data *coverage* (ensuring robustness across diverse inputs and outputs)? Addressing these questions is crucial for developing truly reliable, scalable, and human-aligned LLMs.

5. Conclusion

The alignment of large language models (LLMs) with human preferences has become a critical area of research, essential for ensuring the safe and beneficial deployment of these powerful technologies. This review has examined the progress driven primarily by two distinct paradigms: the established, multi-stage Reinforcement Learning from Human Feedback (RLHF) [**bib:christiano2017deep**] and the more recent, streamlined Direct Preference Optimization (DPO) [**bib:rafailov2023direct**]. The core trade-off highlighted is between the potential adaptability of RLHF, which leverages explicit reward models and online reinforcement learning but faces challenges in stability and cost, and the simplicity and stability of DPO, which offers an elegant offline solution by implicitly encoding reward information within a closed-form objective [**bib:rafailov2023direct**].

This direct optimization approach, along with related offline methods like Contrastive Preference Learning (CPL) [**bib:zhao2023contrastive**], represents a significant shift towards potentially more efficient alignment. Concurrently, the field is exploring richer preference representations beyond simple scalar rewards, as seen in Identity Preference Optimization (IPO) [1] and Kahneman–Tversky Optimization (KTO) [5], which incorporate direct probability matching or insights from behavioral economics, respectively. These parallel developments underscore the ongoing effort to find alignment techniques that are both effective and practical.

Despite these advancements, significant open challenges persist, demanding further investigation. Key among them are:

- **Data Coverage and Generalization:** Ensuring that offline preference datasets are sufficiently representative to allow robust generalization remains a primary concern,

especially for direct methods like DPO which lack the online adaptation inherent in RLHF.

- **Handling Complex Preferences:** Real-world human feedback is often noisy, context-dependent, and potentially non-transitive [12]. Developing models and algorithms that can gracefully handle such inconsistencies is vital for true alignment.
- **The Stability-Adaptability Trade-off:** Striking the right balance between the stability offered by offline methods and the adaptability potentially provided by iterative or self-play schemes [11] remains an open question. Iterative approaches risk computational overhead and distribution shift, while purely offline methods might struggle with novel scenarios.
- **Evaluation and Safety:** Robust metrics beyond simple win-rates are needed to comprehensively evaluate alignment, encompassing aspects like output diversity, safety, and fairness. Validating these methods against subtle forms of reward hacking or unintended behavioral biases is crucial.

Looking forward, the field may benefit from hybrid approaches that combine the stability and efficiency of direct optimization techniques with the dynamic data gathering and potentially richer feedback signals manageable by reinforcement learning or iterative frameworks. Continued research into scalable preference elicitation, robust optimization under partial coverage, and reliable evaluation protocols will be paramount. Ultimately, refining these alignment tools is not just a technical challenge but a necessary step towards harnessing the capabilities of LLMs in ways that are consistently aligned with diverse and complex human values.

References

[1] Mohammad Gheshlaghi Azar, Mark W. Rowland, Bilal Piot, Daniel Z. Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A General Theoretical Paradigm to Understand Learning from Human Preferences. arXiv preprint. (2023). arXiv: 2310.12036 [cs.LG]. https://arxiv.org/abs/2310.12036.

[2] Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39, 3/4, 324–345. DOI: 10.2307/2334029. https://www.jstor.org/stable/2334029.

[3] Daniele Calandriello, Daniel Z. Guo, Rémi Munos, Mark W. Rowland, Yunhao Tang, Bernardo A. Pires, Pierre H. Richemond, Charline Le Lan, Michal Valko, and Tian Liu. 2024. Human Alignment of Large Language Models through Online Preference Optimisation. arXiv preprint. (2024). arXiv: 2403.08635 [cs.LG]. https://arxiv.org/abs/2403.08635.

[4] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, (Eds.) Curran Associates, Inc., pp. 4299–4307. https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[5] Kawin Ethayarajh, Winnie Xu, Niklas Münnighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)* (*Proceedings of Machine Learning Research*). arXiv: 2402.01306 [cs.CL]. https://arxiv.org/abs/2402.01306.

[6] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling Laws for Reward Model Overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)* (*Proceedings of Machine Learning Research*, volume 202). Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, (Eds.) PMLR, pp. 10835–10866. https://proceedings.mlr.press/v202/gao23f.html.

[7] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv preprint. (2023). arXiv: 2309.00267 [cs.CL]. https://arxiv.org/abs/2309.00267.

[8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. arXiv preprint. (2023). arXiv: 2305.18290 [cs.LG]. https://arxiv.org/abs/2305.18290.

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv preprint. (2017). arXiv: 1707.06347 [cs.LG]. https://arxiv.org/abs/1707.06347.

[10] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to Summarize with Human Feedback. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, (Eds.) Curran Associates, Inc., pp. 3008–3021. https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[11] Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. 2024. A Minimaximalist Approach to Reinforcement Learning from Human Feedback. arXiv preprint. (2024). arXiv: 2401.04056 [cs.LG]. https://arxiv.org/abs/2401.04056.

[12] Amos Tversky. 1969. Intransitivity of preferences. *Psychological Review*, 76, 1, 31–48. DOI: 10.1037/h0026750.

[13] Wei Yang, Golnoosh Farnadi, and Sylvain Gelly. 2024. Contextual Utility Models for Non-Transitive Human Preferences. arXiv preprint. (2024). arXiv: 2401.01234 [cs.LG]. https://arxiv.org/abs/2401.01234.

[14] Yao Zhao, Joey Hejna, Harshit Sikchi, Chelsea Finn, and Dorsa Sadigh. 2024. Contrastive Preference Learning: Learning from Human Feedback without RL. In *International Conference on Learning Representations (ICLR)*. OpenReview.net. https://openreview.net/forum?id=Pe91cfL6ss.