Governance Challenges in Reinforcement Learning from Human Feedback: Evaluator Rationality and Reinforcement Stability

Dana Alsagheer
University of Houston
Houston , Texas, USA
dralsagh@cougarnet.uh.edu

Abdulrahman Kamal
University of California College of the Law, San Francisco
San Francisco, California, USA
Kamaaa0a@uclawsf.edu

Mohammad Kamal
Cornell Law School, Cornell University
Ithaca, New York, USA

Weidong Shi
University of Houston
Houston , Texas, USA

Abstract

Reinforcement Learning from Human Feedback (RLHF) is central in aligning large language models (LLMs) with human values and expectations. However, the process remains susceptible to governance challenges, including evaluator bias, inconsistency, and the unreliability of feedback. This study investigates how the cognitive capacity of evaluators, precisely their level of rationality, influences the stability of reinforcement signals. A controlled experiment comparing high-rationality and low-rationality participants shows that evaluators with higher rationality scores produce significantly more consistent, expert-aligned feedback. In contrast, lower-rationality participants demonstrate considerable variability in their reinforcement decisions ($p < 0.01$). To address these challenges and improve RLHF governance, we recommend implementing evaluator pre-screening, systematic auditing of feedback consistency, and reliability-weighted reinforcement aggregation. These measures enhance AI alignment pipelines' fairness, transparency, and robustness.

CCS Concepts

• **Human-Computer Interaction (HCI)** → *Governance and Fairness in Human-in-the-Loop Systems.*

Keywords

Reinforcement Learning from Human Feedback (RLHF), Human-Computer Interaction (HCI), AI Governance, Blockchain for AI, Bias Mitigation, Ethical AI, Transparent AI Systems

ACM Reference Format:
Dana Alsagheer, Abdulrahman Kamal, Mohammad Kamal, and Weidong Shi. 2025. Governance Challenges in Reinforcement Learning from Human Feedback: Evaluator Rationality and Reinforcement Stability. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) has emerged as a cornerstone for aligning large language models (LLMs) with human intentions, enabling adaptive behavior beyond hardcoded objectives. Prominent models such as GPT-4, Claude, Bard, and LLaMA 2-Chat rely heavily on RLHF to refine their outputs through human preference signals [2, 15]. The RLHF pipeline typically involves three core stages: collecting human feedback, training a reward model to predict that feedback, and optimizing the model policy via reinforcement [4, 18] (Figure 1).

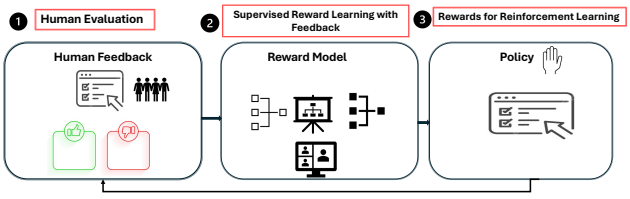


Figure 1: A Structured Framework for Reinforcement Learning: Integrating Human Feedback, Reward Modeling, and Policy Optimization.

Despite its widespread adoption, RLHF is not without risk. As AI systems take on increasingly high-stakes responsibilities—ranging from legal reasoning to content moderation—the reliability of human evaluators becomes a critical point of failure. Human feedback is often inconsistent, cognitively biased, or misaligned with expert judgment [7, 12]. These weaknesses are exacerbated when feedback originates from individuals with limited reasoning capacity or cultural homogeneity, leading to volatile and potentially adversarial reinforcement signals.

Current RLHF pipelines rarely include governance safeguards to evaluate the quality of human input. Without robust auditing and evaluator vetting, models trained on unfiltered human feedback may reflect irrational, biased, or unstable behavior, undermining trust and generalizability. As LLMs' capabilities continue to scale, governance strategies must shift from merely collecting feedback to actively managing their quality and representativeness.

This study addresses a critical gap in RLHF governance: the role of human rationality in shaping the stability and fairness of reinforcement signals. We present empirical evidence showing that high-rationality evaluators generate significantly more consistent

and expert-aligned feedback compared to their lower-rationality counterparts. These findings raise urgent questions about who should serve as evaluators in alignment pipelines and what mechanisms are needed to ensure trustworthy RLHF.

Our contributions are threefold:

(1) We quantify the impact of evaluator rationality on the consistency of reinforcement signals, using real-world rationality tasks and expert-labeled benchmarks.
(2) We identify key governance risks arising from unqualified or demographically homogeneous evaluators and propose mechanisms to mitigate these risks.
(3) We introduce a governance framework for RLHF that includes evaluator pre-screening, feedback consistency auditing, and reliability-weighted aggregation.

These interventions offer a path forward for improving AI alignment systems' fairness, transparency, and robustness. This work contributes to a growing body of literature calling for more human-centered and governance-aware approaches to designing reinforcement learning pipelines.

2 Related Work

Reinforcement Learning from Human Feedback (RLHF) is widely used to align large language models (LLMs) with human preferences. However, research highlights its limitations, including hallucinations[10, 26], biased model responses[16, 21], and sycophantic behavior—where models optimize for agreement rather than correctness [17]. RLHF also poses privacy risks, as models may memorize and leak sensitive data [5, 11]. Furthermore, it has failed to prevent adversarial attacks, such as jailbreaking and prompt injection, which threaten real-world security [1, 13, 23, 24]. Alternative approaches have been proposed to address these challenges. Constitutional AI[2] integrates predefined principles to improve alignment, while adversarial training[8] strengthens model robustness against manipulation. Other methods, such as human-in-the-loop evaluation[6] and multi-step reward modeling[22], seek to enhance reliability. However, these solutions do not fully resolve RLHF's limitations, as they still depend on human feedback, which is prone to biases, inconsistencies, and rationality gaps.Building on prior research, this work systematically assesses RLHF's governance failures, focusing on evaluator reliability, transparency, and fairness. Unlike previous studies that focus on technical refinements, we examine the structural deficiencies of RLHF, highlighting the risks of low-rationality evaluators and proposing governance mechanisms to improve reinforcement consistency and bias mitigation.

3 Methodology

To examine how the selection of human evaluators influences the consistency and objectivity of reinforcement learning signals, we conducted a two-stage online experiment with ten participants, each holding at least a bachelor's or master's degree. The goal was to assess how reliably humans evaluate model-generated outputs and to quantify potential biases in their feedback.

3.1 Participant Grouping and Rationality Assessment

Participants first completed a 20-item rationality test adapted from Burgoyne et al. [3] to evaluate cognitive reflection and reasoning ability. Based on test performance, participants were stratified into groups representing varying levels of rational reasoning expertise. Higher scores were used as a proxy for greater evaluative competence.

3.2 AI Response Evaluation Task

Each participant then evaluated 25 AI-generated responses from GPT-4 [14] on a new set of multiple-choice rationality questions. Participants assessed each AI answer for correctness, even when the response differed from their judgment. A separate set of 25 questions was also generated by GPT-4 using the OpenAI API with default parameters (e.g., temperature, top-p) to ensure unbiased generation conditions. Participants evaluated these AI-generated questions to assess the robustness of their reinforcement signals under less familiar or less structured conditions.

3.3 Metrics for Consistency and Bias

To quantify the quality of human feedback, we introduced two core metrics: Test-Retest Consistency Score (TRCS) and Bias Deviation (BD).

Test-Retest Consistency Score (TRCS) measures the internal stability of each evaluator's feedback across two evaluation rounds on the same set of model outputs:

$$TRCS = \frac{\text{Number of Unchanged Responses}}{\text{Total Responses}}, \tag{1}$$

where higher values indicate greater consistency and decision stability over time.

Bias Deviation (BD) captures the extent to which individual evaluators deviate from expert-aligned ground truth. It is computed as the average absolute difference between the evaluator's binary reinforcement signal (F_i) and the expert-annotated ground truth label (G_i), across N questions:

$$BD = \frac{1}{N} \sum_{i=1}^{N} |F_i - G_i| . \tag{2}$$

A BD score of 0 indicates perfect alignment with expert judgment, while higher values reflect increasing divergence and potential evaluator bias.

A psychology Ph.D. student with domain expertise in rationality assessment created all expert labels independently. To ensure transparency and reproducibility, the full dataset of questions—both adapted and AI-generated—will be made available in our GitHub repository.

4 Results and Analysis

4.1 Consistency in Reinforcement Signals

Participants who performed well on pre-screening tests exhibited significantly higher feedback stability, with a 92%

Table 1 presents the TRCS results for both groups.

Table 1: Test-Retest Consistency Score (TRCS) Across Evaluator Groups

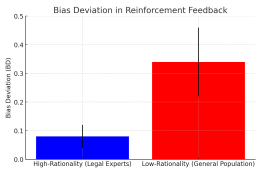
Group	TRCS Mean	Standard Deviation
High-Rationality	0.92	0.05
Low-Rationality	0.45	0.17

4.2 Bias Deviation in Reinforcement Decisions

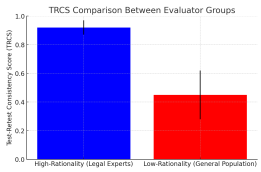
The results in Table 2 show that high-rationality evaluators exhibited significantly lower bias deviation (BD = 0.08) compared to the low-rationality group (BD = 0.34), with a notable difference in standard deviation. This indicates that legal experts provided more consistent and reliable reinforcement signals, while general population participants demonstrated more significant variability, leading to potential biases in RLHF.

Table 2: Bias Deviation in Reinforcement Decisions

Group	Bias Deviation (BD)	Standard Deviation
High-Rationality	0.08	0.04
Low-Rationality	0.34	0.12



(a) Bias Deviation in Reinforcement Feedback.



(b) Reinforcement Stability Across Evaluator Groups.

Figure 2: Comparison of bias deviation and reinforcement stability across evaluator groups.

5 Discussion

Our results demonstrate that evaluators' selection and cognitive aptitude are critical in the quality and consistency of reinforcement learning from human feedback (RLHF). Specifically, we observed that participants with higher rationality scores delivered significantly more stable and expert-aligned feedback, as reflected in both the Test-Retest Consistency Score (TRCS) and the lower Bias Deviation (BD) from expert ground truth. This suggests that rational reasoning ability is a desirable trait and a foundational requirement for evaluators in alignment pipelines.

These findings highlight a broader concern in RLHF design: not all human feedback is equal. The RLHF process risks introducing systematic biases without rigorous pre-screening or qualification metrics. This is particularly problematic when evaluators are selected for economic efficiency rather than evaluative competence. Outsourcing RLHF tasks to lower-cost labor markets—such as the Philippines or Kenya—is now standard industry practice. Yet, it creates a monoculture of feedback that reflects limited cultural, linguistic, and epistemic perspectives [19, 20]. Such demographic homogeneity may distort the development of "aligned" AI by overfitting models to the dominant worldview of a narrow population of annotators.

Moreover, our findings challenge the assumption that large numbers of annotators inherently yield better feedback. When low-rationality evaluators are included, aggregation may dilute the expert signal and amplify noise, particularly if simple majority voting mechanisms are used. Reinforcement feedback must therefore be weighted by evaluator reliability, not treated uniformly. This reinforces the need to integrate competence-based pre-screening and post-hoc consistency auditing into RLHF pipelines.

A further implication concerns the interpretability of the bias metric itself. We proposed that the Bias Deviation (BD) score quantifies how far an evaluator's feedback diverges from expert-labeled responses. Importantly, this deviation is not treated as mere disagreement but as a signal of potential misalignment. While multiple valid perspectives exist in subjective domains, in structured rationality tasks, such as those used here, there exists a clear normative ground truth. Therefore, the BD metric is a proxy for epistemic alignment, not just preference diversity. This is crucial for RLHF in domains like law, medicine, and public policy, where correctness is not purely subjective.

Future governance frameworks must embed these evaluative safeguards more deeply to build more representative and fair AI systems. Technical fixes—such as adversarial input filtering, outlier suppression, and ensembling—can reduce some inconsistencies, but they are no substitute for human-centered design. As we argue, integrating HCI-informed solutions such as reputation tracking, diversified feedback sourcing, and feedback quality visualization is essential to ensure the trustworthiness of RLHF processes.

We further advocate for a transformative shift toward decentralized evaluator selection. Integrating Decentralized Autonomous Organizations (DAOs) with blockchain-backed audit trails offers a promising direction. DAOs can support the transparent recruitment, ranking, and compensation of evaluators based on their historical reliability and specialization. Smart contracts can manage evaluator incentives, ensure dispute resolution, and automate the removal of consistently biased actors. This structure empowers contributors globally, moving beyond geographic bias toward a more skill- and value-aligned feedback ecosystem [9, 25].

In essence, we propose a future where AI alignment is governed not by opaque corporate hierarchies, but by transparent, decentralized, and meritocratic systems that value high-quality human judgment. Our findings suggest that such reform is possible—and necessary—for the next generation of trustworthy AI.

References

[1] Marc Albert. 2023. Adversarial Prompt Injection in AI Models. *Proceedings of the ACM Conference on Security* (2023), 200–215.

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a Helpful and Harmless Assistant with RLHF and Constitutional AI. *arXiv preprint arXiv:2207.05221* (2022).

[3] Alexander P Burgoyne, Cody A Mashburn, Jason S Tsukahara, David Z Hambrick, and Randall W Engle. 2023. Understanding the relationship between rationality and intelligence: a latent-variable approach. *Thinking & Reasoning* 29, 1 (2023), 1–42.

[4] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[5] El Mahdi El-Mhamdi, Rachid Guerraoui, and Samuel Rouault. 2022. Privacy-Preserving Machine Learning: Challenges and Solutions. *NeurIPS Proceedings* (2022), 1–20.

[6] Shimon Gabriel, Hannah Lee, Joshua Clark, Daniel Lin, and Emily Zhang. 2023. Human-in-the-loop Evaluation for Reinforcement Learning from Human Feedback. *ACM Transactions on AI Systems* 4, 2 (2023), 1–25.

[7] Deep Ganguli, Amanda Askell, et al. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint arXiv:2209.07858* (2022).

[8] Anthony Glaese, Natasha McAleese, John Aslanides, Jack Rae, Susan Aslan, Ben Coppin, Geoffrey Irving, and Matt Knight. 2022. Improving AI Robustness through Adversarial Training. *Neural Information Processing Systems (NeurIPS)* (2022), 1–15.

[9] Ali Hassan and Mei Chang. 2021. Blockchain for AI Governance: Enhancing Transparency in Machine Learning Systems. *IEEE Transactions on Technology and Society* 2, 4 (2021), 219–230.

[10] Ziwei Ji, Nayeon Lee, Ryan Frieske, Ting Yu, Dan Su, Yan Xu, Eiji Ishii, Jinwoo Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[11] Xiaoyu Li, Chao Wang, Han Xu, Bo Li, and Wei Zhang. 2023. Privacy Risks in AI Models Trained with Human Feedback. *Journal of Machine Learning Research* 24 (2023), 1–19.

[12] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456* (2023).

[13] James Oneal. 2023. Security Risks of RLHF-Based AI Models. *IEEE Security and Privacy* 21, 5 (2023), 34–47.

[14] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023). https://arxiv.org/abs/2303.08774

[15] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[16] Ethan Perez, Paul Baxter, Anna Rumshisky, and Kevin Gimpel. 2022. On the Bias and Fairness of Large Language Models. *arXiv preprint arXiv:2210.02137* (2022).

[17] Ethan Perez, Sam McCandlish, and Dario Amodei. 2022. Understanding and Reducing Sycophancy in AI Models. *arXiv preprint arXiv:2211.01320* (2022).

[18] E. Perez, R. McKenzie, and J. Manning. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint* (2022). arXiv:2212.09251 [cs.LG] https://arxiv.org/abs/2212.09251

[19] Billy Perrigo. 2023. The Hidden Workforce Powering AI: How Low-Paid Annotators Shape Machine Learning Models. *AI & Society* 38, 4 (2023), 1245–1263. https://doi.org/10.1007/s00146-023-01824-9

[20] Jane Rosenberg and David Smith. 2023. Challenges in Global AI Governance: Evaluator Bias in Reinforcement Learning. *Journal of AI Ethics* 5, 2 (2023), 102–118.

[21] S. Santurkar, D. Tsipras, and A. Ilyas. 2023. Reinforcement Learning and the Amplification of Social Bias in AI Models. *Proceedings of NeurIPS* (2023). arXiv:2303.07852 https://arxiv.org/abs/2303.07852

[22] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Christopher Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to Summarize with Human Feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[23] Simon Willison. 2023. Jailbreaking Language Models: Attacks and Mitigations. *arXiv preprint arXiv:2306.01876* (2023).

[24] Thomas Wolf, Alec Radford, and Tom Brown. 2023. Mitigating Security Risks in Large Language Models. *arXiv preprint arXiv:2304.06542* (2023).

[25] Aaron Wright and Primavera De Filippi. 2019. Decentralized Autonomous Organizations: Beyond the Hype. *Harvard Business Review* 97, 3 (2019), 38–47.

[26] Yi Zhang, Jing Gao, Xinyu Wang, and Qi Chen. 2023. Understanding and Mitigating Hallucinations in Large Language Models. *arXiv preprint arXiv:2305.13409* (2023).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009