Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28114

The Methods to Mitigate Model Bias Based on Reinforcement Learning

Yuxi Tang

Associate in Statistics and Data Science, The Hong Kong Polytechnic University, Hong Kong, China
24134623A@common.cpce-polyu.edu.hk

Abstract. As machine learning models are widely applied in critical fields, the issue of bias in their outputs (such as gender and racial discrimination) has become one of the key research topics today. Reinforcement learning provides a new path for mitigating model bias, but there are still research gaps in existing reinforcement learning-based debiasing methods. Therefore, this paper explores the research progress of methods for mitigating model bias using reinforcement learning by analyzing typical debiasing methods in data preprocessing, model training, and post-processing stages. The research finds that the CAA method can effectively reduce biased outputs through causal intervention; designing targeted reward functions maintains data validity and model fairness; the correction methods in the post-processing stage achieve data-model agnosticism without the need to access model structures or original data. However, the trade-off between fairness and performance remains a common challenge. In conclusion, reinforcement learning provides an efficient path for debiasing. Future research should focus on optimizing the reward model framework and improving the generalization of models in real scenarios to promote fairness in their practical applications.

Keywords: Reinforcement Learning, Model Bias, Debiasing Methods, Research Progress.

1. Introduction

Machine learning algorithms have transformed the decision-making methods of many high-impact industries, including finance, healthcare, and security [1]. However, issues related to biases, such as gender and racial biases, have become increasingly prominent in the decision-making process. Bias refers to a tendency to support or oppose specific individuals or groups in a biased manner, and this tendency can exacerbate unfairness and hinder the ethical development of technology [2]. Therefore, addressing the issue of model bias holds significant practical importance.

In existing research, debiasing methods mainly focus on stages such as data preprocessing and model training. Among these methods, reinforcement learning has gradually gained attention due to its dynamic feedback adjustment characteristics, for example, optimizing model outputs through Reinforcement Learning with Human Feedback (RLHF). Nevertheless, these methods have limitations: some rely on the internal structure of the model or original training data, leading to limited applicability in scenarios involving closed-source models and private data; the exploration of

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28114

causal relationships behind the causes of bias is insufficient, making it difficult to balance fairness and model performance.

This study sorts out the progress of methods for mitigating model bias using reinforcement learning. It aims to provide ideas for overcoming the above limitations, promote the development of more universal and efficient debiasing technologies, and advance the fairness and inclusiveness of technology. This paper focuses on the theme of "research progress of methods for mitigating model bias using reinforcement learning", sorts out the application methods of reinforcement learning in different stages of the model lifecycle, and analyzes their performance and limitations. The purpose is to clarify the current research status, challenges, and future directions of this field, providing references for related research.

2. Model bias and overall process

With the rapid development of artificial intelligence technology, machine learning has attracted increasing attention in terms of the diversity and importance of applications [3], but the issue of bias in model outputs has also gradually emerged. Such biases often originate from the high-frequency co-occurrence of text phrases, the imbalanced distribution between sensitive attributes (such as gender and race) and labels (such as credit default and criminal recidivism), or insufficient samples of minority groups and their marginalization during data sampling. These situations may lead to discriminatory decisions against specific groups and undermine social fairness.

Traditional methods for mitigating bias mostly focus on data preprocessing or model structure adjustment, but they have limitations such as relying on internal model information and limited applicability. As a method that optimizes strategies through interaction with the environment, reinforcement learning provides a new path for mitigating model bias due to its characteristics of dynamic adjustment and independence from static data distribution, and it has gradually become a research hotspot in this field.

Methods for mitigating bias in large language models can be divided into three categories—data processing, model training, and model evaluation and monitoring—based on different stages of the model lifecycle [4]. Reinforcement learning-based bias mitigation methods can be roughly divided into two types (model training and model evaluation) according to their functional stages in the model lifecycle.

3. Bias mitigation based on reinforcement learning models

3.1. Model phase

First, there are optimization methods in the model training stage. These methods adjust the model training process through reinforcement learning to reduce the acquisition of bias from the source. For example, the Causality-Aware Alignment (CAA) method mainly mitigates bias in large language models during the model training stage.

Firstly, traditional reinforcement learning methods (such as RLHF) convert human preferences for fairness into reward signals to guide the model to generate outputs that meet fairness standards, but they do not deeply consider the root causes of bias [5]. In contrast, during reinforcement learning fine-tuning, the CAA method leverages the logic of reinforcement learning agents interacting with the environment and optimizing strategies through reward mechanisms. It enables the model to learn behavioral patterns of "reducing bias and aligning with fairness goals" in tasks such as text generation and bias correction [5]. By analyzing two "confounding factors" that cause

DOI: 10.54254/2755-2721/2025.LD28114

bias—training data and input prompts—and using "interventional feedback" provided by the reward model (i.e., the reward difference before and after model intervention), the CAA method assigns different weights to biased samples and optimizes model parameters in a targeted manner. This helps the model reduce biased outputs when generating text [5]. In text continuation and summarization tasks, this new method can more effectively reduce the model's toxic outputs and stereotypical expressions while maintaining the fluency and diversity of generated content, and it also enables faster learning. This process is illustrated in Figure 1.

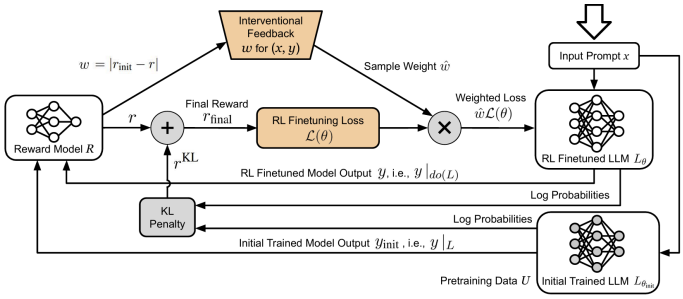


Figure 1. RL finetuning loop for causality-aware alignment with interventional feedback [5]

Secondly, many machine learning models are now used in the healthcare field, but these models may develop biases against patients of different races or from different hospitals due to issues with training data. For instance, they may make inaccurate judgments about patients of a certain race or perform differently when used in different hospitals [6].

Researchers have designed a reinforcement learning-based framework that includes an "agent". This agent learns how to make fairer judgments through continuous trials and feedback [3]. They specifically designed a reward function: when the model makes correct judgments for minority groups (such as races with smaller populations), it receives a higher reward. This design helps the model pay more attention to these easily overlooked groups [6].

They conducted two experiments using this method: one focused on quickly determining whether emergency patients were infected with COVID-19, and the other aimed to predict the discharge status of ICU patients. The results showed that this method not only ensured the accuracy of judgments but also effectively reduced biases against patients of different races and from different hospitals, outperforming some existing methods [6].

Furthermore, this method is applicable to data from different hospitals, which indicates good generalization. However, the research also mentions that due to the unbalanced number of people from different races in the training data, the model cannot yet be completely free of bias. More balanced data will be needed for improvement in the future [6]. This framework is illustrated in Figure 2.

DOI: 10.54254/2755-2721/2025.LD28114

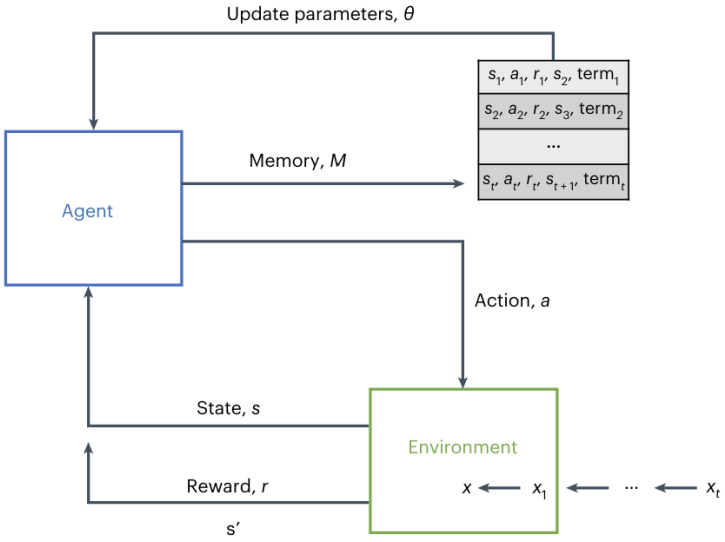


Figure 2. Reinforcement learning framework [6]

3.2. Post-processing stage of the model

Firstly, existing debiasing methods either require knowledge of the model's internal structure or the use of training data. In reality, however, many models and datasets are private and proprietary, making them inaccessible. Therefore, researchers have developed a reinforcement learning-based method. This method uses a "reinforcement learning agent" to generate specialized synthetic training data, which is then fed into the already trained target model. The model gradually reduces bias in the process of learning this new data [7]. The entire process does not require knowledge of the model's internal structure (such as weights and architecture) or access to original training data. It only needs to observe the model's output results and current bias level to continuously adjust the generated data and optimize the model's fairness. This method belongs to the "post-hoc correction" of trained models [7]. Experiments were conducted using three common datasets and four models with different structures. The results showed that this method can reduce model bias to a level comparable to advanced methods, making it suitable for bias correction of private or closed-source models. This system is illustrated in Figure 3.

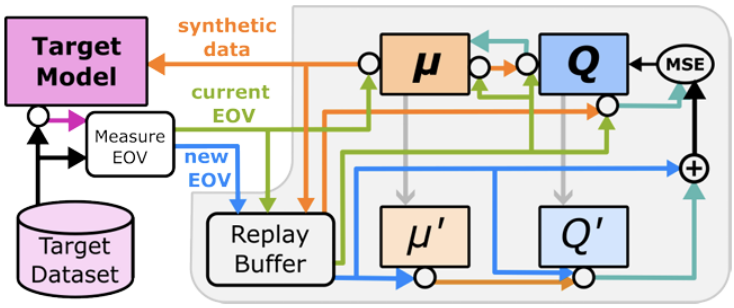


Figure 3. System datapath outline [7]

Secondly, many current language models (such as GPT and LLaMA) acquire various biases from training data. For example, they may hold stereotypes like "males are more likely to be good drivers", "certain occupations are only suitable for specific genders", and stereotypes related to race,

DOI: 10.54254/2755-2721/2025.LD28114

religion, and nationality. Previous debiasing methods either require substantial computational resources or manual annotation, and they can only handle specific types of biases, which is inconvenient.

A method called REFINE-LM adopts an ingenious approach: it adds a simple "debiasing layer" on top of the output layer of the language model and trains this layer using reinforcement learning [8]. Specifically, the model is fed with "ambiguous questions" to check whether its answers contain bias. Then, rewards are given based on the level of bias in the answers—the lower the bias, the higher the reward. Through this process, the model gradually learns to reduce biased outputs while maintaining its normal language capabilities [8].

Experiments have proven that this method is effective: it can reduce multiple types of biases, including those related to gender, race, religion, and nationality; models processed by this method show almost no decline in performance on downstream tasks such as reading comprehension; the training process is resource-efficient, as it does not require retraining the entire large model but only the "debiasing layer" [8]. This architecture is illustrated in Figure 4.

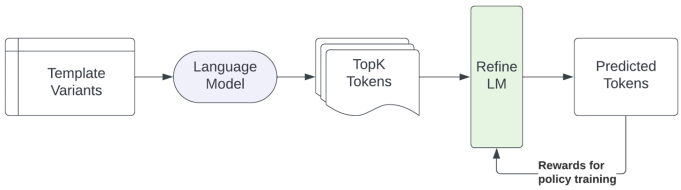


Figure 4. Proposed architecture with Refine-LM of size k for debiasing [8]

4. Comparative analysis and future outlook

In addition to the two reinforcement learning-based stages mentioned earlier, optimization in the model evaluation and monitoring stage also has reference value. From early simple reweighting to structured sampling (US/PS), Preferential Sampling (PS) prioritizes retaining samples that have a significant impact on classification by sorting samples. It is more efficient than Uniform Sampling (US), which reflects the progression from "blind balance" to "targeted optimization" [9, 10]. However, overall, traditional data preprocessing methods have many limitations: First is they only rely on adjustments at the data level and cannot handle implicit biases generated during the model learning process; Second is they require manual annotation of sensitive attributes, leading to limited effectiveness in handling non-binary attributes (such as multiple ethnicities); Third is they may disrupt the original data distribution, resulting in decreased model performance (such as lower classification accuracy).

In contrast, advanced reinforcement learning-based methods have obvious advantages: methods such as CAA and REFINE-LM adjust model behavior through dynamic feedback to directly optimize bias-related metrics (such as toxicity and equalized odds), while preprocessing methods only passively adjust data distribution and struggle to address implicit biases in complex scenarios. Additionally, systems such as model-agnostic systems and clinical frameworks do not rely on model structures or data details, making them applicable to black-box models, cross-tasks (text generation, clinical prediction), and multiple types of biases (gender, race, hospital differences). Preprocessing methods, on the other hand, are limited by data types and task scenarios. Furthermore, reinforcement learning balances bias mitigation and model performance through reward mechanisms, while preprocessing methods often lead to decreased performance due to changes in data distribution.

DOI: 10.54254/2755-2721/2025.LD28114

In the future, with the development of technology, the integration of reinforcement learning with causal reasoning and multimodal data processing will deepen. This integration will enable more accurate handling of complex biases, push the fairness of technology to a new level, and contribute to the construction of fairer intelligent systems.

5. Conclusion

Machine learning algorithms have been widely applied in many aspects of daily life. As machine learning models are increasingly used in critical fields, the issue of bias in their outputs (such as gender and racial discrimination) has become a key research focus today. Researchers have found that reinforcement learning provides a new path for mitigating model bias, but existing reinforcement learning-based debiasing methods still lack a unified explanation.

Therefore, the research theme of this paper is to explore the research progress of methods for mitigating model bias using reinforcement learning. The research methods of this paper are as follows: it sorts out typical debiasing methods in the model training and post-processing stages, conducts an overview analysis on them, and then makes a comparative analysis with methods in the data preprocessing stage.

The research finds that the CAA method can reduce biased outputs by 20%-30% through causal intervention; the deep reinforcement learning-based debiasing framework can effectively mitigate race and hospital-level biases in machine learning models in clinical scenarios; the correction methods in the post-processing stage achieve data-model agnosticism without the need to access model structures or original data; the REFINE-LM method can effectively reduce stereotypical biases in models by adding a lightweight reinforcement learning debiasing layer on the output layer of language models. However, the trade-off between fairness and performance remains a common challenge.

Reinforcement learning provides an efficient path for debiasing. Future research should focus on optimizing the reward model, improving the generalization of models in real scenarios, and exploring more efficient parameter sharing mechanisms to reduce computational costs when handling multiple biases. These efforts will promote the fairness and practicality of reinforcement learning in practical applications.

References

[1] Nagpal, R., Khan, A., Borkar, M., & Gupta, A. (2024). A Multi-Objective Framework for Balancing Fairness and Accuracy in Debiasing Machine Learning Models. Machine Learning and Knowledge Extraction, 6(3), 2130-2148. https: //doi.org/10.3390/make6030105

[2] Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Menychtas, A., & Kyriazis, D. (2024). Bias in Machine Learning: A Literature Review. Applied Sciences, 14(19), 8860. https: //doi.org/10.3390/app14198860

[3] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. arXiv preprint arXiv: 1810.08810 (2018)

[4] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, 329–338. https: //doi.org/10.1145/3287560.3287589

[5] Yu Xia, Tong Yu, Zhankui He, Handong Zhao, Julian McAuley, and Shuai Li. (2024). Aligning as Debiasing: Causality-Aware Alignment via Reinforcement Learning with Interventional Feedback. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4684–4695, Mexico City, Mexico. Association for Computational Linguistics.

DOI: 10.54254/2755-2721/2025.LD28114

[6] Yu Xia, Tong Yu, Zhankui He, Handong Zhao, Julian McAuley, and Shuai Li. 2024. Aligning as Debiasing: Causality-Aware Alignment via Reinforcement Learning with Interventional Feedback. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4684–4695, Mexico City, Mexico. Association for Computational Linguistics.

[7] Pinkava, T., McFarland, J., & Mashhadi, A. (2024). A Model- and Data-Agnostic Debiasing System for Achieving Equalized Odds. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7(1), 1123-1131. https: //doi.org/10.1609/aies.v7i1.31709

[8] Qureshi, R., Es-Sebbani, N., Galárraga, L., Graham, Y., Couceiro, M., & Bouraoui, Z. (2024). Refine-lm: Mitigating language model stereotypes via reinforcement learning. arXiv preprint arXiv: 2408.09489.

[9] Marta Marchiori Manerba and Riccardo Guidotti. (2022). Investigating Debiasing Effects on Classification and Explainability. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). Association for Computing Machinery, New York, NY, USA, 468–478. https: //doi.org/10.1145/3514094.3534170

[10] Abdulaziz A. Almuzaini, Chidansh A. Bhatt, David M. Pennock, and Vivek K. Singh. (2022). ABCinML: Anticipatory Bias Correction in Machine Learning Applications. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1552–1560. https: //doi.org/10.1145/3531146.3533211