Personalizing Fairness: Adaptive RL with User Diversity Preference for Recommender Systems

Luana G. B. Martins[1], **Bryan L. M. de Oliveira** [1, 2], **Bruno Brandão**[1, 2], **Telma W. de L. Soares**[1, 2], **Marlesson R. O. Santana**
luanagbmartins@gmail.com

[1]**Advanced Knowledge Center for Immersive Technologies – AKCIT, Brazil**
[2]**Institute of Informatics, Federal University of Goiás, Goiânia, Brazil**

Abstract

Reinforcement learning is increasingly applied to optimize recommender systems for long-term user engagement and system objectives. However, a significant challenge remains in ensuring fair supplier exposure alongside user relevance, as traditional methods often lead to popularity bias. Addressing this challenge by adaptively balancing relevance and fairness can lead to more sustainable, equitable digital platforms and improved long-term user engagement. We introduce A2Fair, a RL framework that personalizes recommendations by dynamically balancing relevance and exposure fairness through an adaptive reward function that considers individual user diversity preferences and a rich state representation.

1 Introduction

Recommender systems (RS) in digital marketplaces personalize user experience and drive engagement. However, established RS methodologies like collaborative and content-based filtering (Das et al., 2007; Marlin & Zemel, 2004; Su & Khoshgoftaar, 2009; Yang et al., 2014; Adomavicius & Tuzhilin, 2005; Kompan & Bieliková, 2010; Phelan et al., 2011) can be limited in dynamic environments with long-term interaction effects. Reinforcement learning (RL) offers a robust alternative for optimizing decision policies by learning from long-term environmental interactions. Consequently, RL-based recommender systems (RLRS) are developed to enhance personalization and efficiency in these settings (Dulac-Arnold et al., 2015; Zhao et al., 2018; Zheng et al., 2018; Zou et al., 2019; Liu et al., 2020b).

A key challenge in marketplace system is ensuring fair supplier exposure alongside user-centric metrics. Conventional RS and naive RL implementations can create popularity bias, where few items or suppliers dominate recommendations. This "rich-get-richer" phenomenon (Mehrotra et al., 2018; Singh & Joachims, 2018) hinders new or niche suppliers, limits user discovery, and impacts marketplace health and diversity (Abdollahpouri et al., 2020; Pitoura et al., 2021; Wang et al., 2023). Addressing this requires balancing immediate user relevance with supplier exposure (Abdollahpouri et al., 2020), considering the dynamic, sequential nature of user interactions and evolving preferences (Liu et al., 2019; D'Amour et al., 2020; Creager et al., 2020b; Chen et al., 2023; Gohar et al., 2025; Tang et al., 2025).

Users exhibit diverse preferences for content variety; some prefer specific supplier groups, while others welcome broader recommendations (Mehrotra et al., 2018; Mansoury et al., 2023; Prent & Mansoury, 2025), suggesting that a one-size-fits-all approach to fairness and relevance is suboptimal. Our approach addresses this by dynamically balancing recommendation relevance with supplier exposure fairness, considering individual user affinity for diversity. We use RL to learn

a policy that personalizes for both relevance and fairness by modeling sequential interactions as a markov decision process (MDP) and designing a reward function that incorporates user diversity preference.

This approach yields RL agents for RS that are effective by traditional relevance metrics, robust to fairness concerns, and adaptable to diverse user behaviors. Where dynamic adaptation to user preferences and system-level objectives like fairness is paramount, RL can foster sustainable and equitable digital platforms. Successfully balancing accuracy and fairness through such an RL framework can improve long-term user engagement and cultivate a healthier, more diverse marketplace ecosystem.

This work contributes A2Fair, an RL-based solution for balancing relevance and fairness in recommender systems. Experiments on public datasets demonstrate that our approach, which explicitly designs the RL agent to consider user affinity for diversity, improves both recommendation relevance and exposure fairness.

2 Fairness of Exposure in Recommender Systems

Recommender systems personalize suggestions based on user preferences, which can inadvertently lead to unequal exposure among items. This disparity occurs because optimizing solely for user-centric metrics often neglects broader system-level objectives, such as ensuring fair visibility for item providers (Abdollahpouri et al., 2020). Although traditional RS evaluations prioritize user-focused metrics like precision, recall, and diversity (Abdollahpouri et al., 2020), incorporating provider perspectives is crucial for a balanced ecosystem. In multilateral recommendation environments, establishing an equilibrium between relevant recommendations for users and equitable exposure for providers is a fundamental challenge for ecosystem sustainability.

The dynamic nature of recommender systems, where user interactions continuously reshape the environment, introduces additional complexity. Applying static fairness criteria at each interaction step can, paradoxically, worsen long-term unfairness (Creager et al., 2020a; D'Amour et al., 2020; Zhang et al., 2020; Deldjoo et al., 2023; Wang et al., 2023; Mansoury & Mobasher, 2023; Gohar et al., 2025; Tang et al., 2025). Several RL-based approaches have been proposed to address these dynamic challenges. For example, Wen et al. (2021) formalize fair decision-making within MDPs to achieve long-term group fairness. Yadav et al. (2021) integrate meritocratic fairness principles with policy gradients. Other strategies involve categorizing items by popularity and applying RL with fairness constraints (Ge et al., 2021), or employing demographic parity where deviations from fair exposure serve as learning constraints for policy gradients (Singh & Joachims, 2019). Alternative frameworks, such as multi-objective RL (Ge et al., 2022) and fairness-aware multi-armed bandits (Joseph et al., 2016; Metevier et al., 2019; Patil et al., 2021; Wang et al., 2021), also provide solutions for balancing the exploration-exploitation trade-off alongside fairness objectives.

Items can be categorized into groups based on various criteria, such as sensitive attributes or platform monetization strategies. This grouping facilitates the management of item exposure, enabling either equal or strategically differentiated visibility across groups. Consequently, platform visibility can be tailored to group-specific exposure targets, which are often defined by importance weights reflecting platform strategies. Formally, let $\mathcal{A}$ be the set of all items, partitioned into l distinct groups, $G = \{g_1, \ldots, g_l\}$. For each group $g_i \in G$, $\mathcal{A}_{g_i} \subseteq \mathcal{A}$ denotes the subset of items belonging to that group. Group exposure fairness (Singh & Joachims, 2018; Liu et al., 2020c) is quantified using these item groupings. The exposure of a group g_i up to time t, denoted as x_t^i, is the i-th component of the exposure distribution vector $x_t \in \mathbb{R}_+^l$ across all l groups:

$$x_t^i = \frac{\sum_{k=1}^{t} \mathbf{1}_{\mathcal{A}_{g_i}}(a_k)}{\sum_{i'=1}^{l} \sum_{k=1}^{t} \mathbf{1}_{\mathcal{A}_{g_{i'}}}(a_k)}, \tag{1}$$

where $\mathbf{1}_{\mathcal{A}_{g_i}}(a_k)$ is the indicator function, which equals 1 if item a_k belongs to group $\mathcal{A}_{g_i}$, and 0 otherwise. Weighted proportional fairness aims to maximize $PropFair$:

$$PropFair = \sum_{i=1}^{l} w_i \log(1 + x_t^i), \tag{2}$$

subject to $\sum_{i=1}^{l} x_t^i = 1$ and $x_t^i \geq 0$. Here, $w_i \in \mathbb{R}_+$ are predefined importance weights for each group. The optimal exposure distribution x_*^i for group g_i under this metric is: $x_*^i = \frac{w_i}{\sum_{i'=1}^{l} w_{i'}}$

Weighted proportional fairness is one of several fairness definitions; the choice of metric is highly context-dependent, and no single definition is universally applicable. Liu et al. (2020c), for example, employed this fairness definition within an RL framework. They argued that only post-exposure actions with direct commercial value (e.g., clicks) should contribute to exposure calculations, thereby disregarding items that are shown but do not lead to immediate conversion. However, this perspective overlooks the potential of initial exposure—even without immediate conversion—to act as a precursor to future engagement, such as later conversions, product discovery, or increased familiarity. Therefore, an exposure that does not yield a high immediate reward might still be instrumental in a sequence of interactions that ultimately leads to a higher cumulative reward in the long term.

3 Combining relevance and fairness

We model the recommendation process as a MDP, where the RS acts as the agent and user interactions constitute the environment. At each discrete time step t, the agent observes the current state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$ (recommends an item), receives user feedback y_{a_t} (e.g., a click or skip), and obtains a reward r_t. The agent's objective is to learn an optimal policy π^* that maximizes the cumulative reward over time. The core components of this MDP are defined as follows:

The immediate relevance reward, $R_{\text{relevance}}(u, a_t)$, for recommending a_t to user u is based on feedback y_{a_t}. We set $R_{\text{relevance}}(u, a_t) = +1$ for positive interactions (e.g., clicks, purchases), signifying high relevance, and $R_{\text{relevance}}(u, a_t) = -1$ otherwise.

To promote fair exposure for item providers, items are categorized into l distinct groups, $G = \{g_1, \ldots, g_l\}$, with each item belonging to exactly one group $\mathcal{A}_{g_i} \subseteq \mathcal{A}$. These groups are typically defined by system designers based on criteria such as shared item attributes, provider identity, or platform-specific policies. The exposure of a group g_i at time t, denoted x_t^i (Eq. 1), is considered fair when it aligns with a predetermined ideal exposure x_*^i, function of the group's importance weight w_i. This framework empowers recommender system developers to adjust group exposure in line with platform guidelines, thereby balancing strategic objectives with the goal of fair exposure.

The fairness component of the reward, $R_{\text{fairness}}(a_t)$, is designed to guide the agent towards achieving the target exposure distribution. It is calculated based on the disparity between the current exposure x_t^i (Eq. 1) of the group to which item a_t belongs and its optimal exposure x_*^i:

$$R_{\text{fairness}}(a_t) = \sum_{i=1}^{l} \mathbf{1}_{\mathcal{A}_{g_i}}(a_t)(x_*^i - x_t^i), \tag{3}$$

The term $(x_*^i - x_t^i)$ quantifies the fairness gap for the selected item's group: a positive value indicates that the group is under-represented, while a negative value signifies over-representation.

Users exhibit significant variability in their sensitivity to content diversity: some users consistently prefer items from specific groups, while others are more open to recommendations from a broader range of categories (Mehrotra et al., 2018; Mansoury et al., 2023; Prent & Mansoury, 2025). To capture this, we use user u's history of positive interactions $H_u = \{e_{a_1}, \ldots, e_{a_m}\}$, where m is the total count of relevant items. Each item $a \in H_u$ is associated with an embedding e_a, which captures its characteristics.

The challenge lies in quantifying a user's preference for items similar to those they previously favored. We use a diversity coefficient based on the cosine similarities between the embeddings of items in H_u. The preference for diverse content is then defined as:

$$\eta(u) = 1 - \mu_{\zeta_u}, \tag{4}$$

where μ_{ζ_u} is the mean cosine similarity of items in H_u, and $\eta(u) \in [0, 1]$. $\eta(u) \approx 0$ suggests preference for similar items; $\eta(u) \approx 1$ indicates preference for variety.

We propose an adaptive reward function to optimize RS, dynamically balancing relevance for users with low diversity affinity and fairness for users with high diversity affinity. It integrates relevance and fairness using a dynamic weighting factor:

$$R(u, a_t) = ((1 - \eta(u)) * R_{\text{relevance}}(u, a_t)) + (\eta(u) * R_{\text{fairness}}(a_t)). \tag{5}$$

As a user's diversity preference $\eta(u)$ (Eq. 4) increases, the system emphasizes fairness. Conversely, as $\eta(u)$ decreases, the focus shifts to relevance. This approach aims for a fairer recommender system that adapts to individual user preferences while promoting equitable group visibility. This dynamic balance enables satisfying immediate user needs while promoting long-term fairness and diversity.

4 Proposed framework

We present A2Fair, a framework for fair recommendations considering user profile, using an Actor-Critic algorithm with three main components: state representation module, actor, and critic. The state representation module (Figure 1) integrates user and system context, defined by four pillars:

1. User Understanding: Integrating user representation allows considering unique characteristics for more relevant recommendations.
2. Recent Preferences: Last N positively interacted items reflect current preference trends
3. Detailed Interaction Modeling: Explicitly incorporating user-item interactions enhances RL capabilities for identifying preference patterns.
4. Fairness and Diversity Promotion: Including group exposure distribution clarifies current fairness balance, allowing strategic prioritization of underrepresented groups.

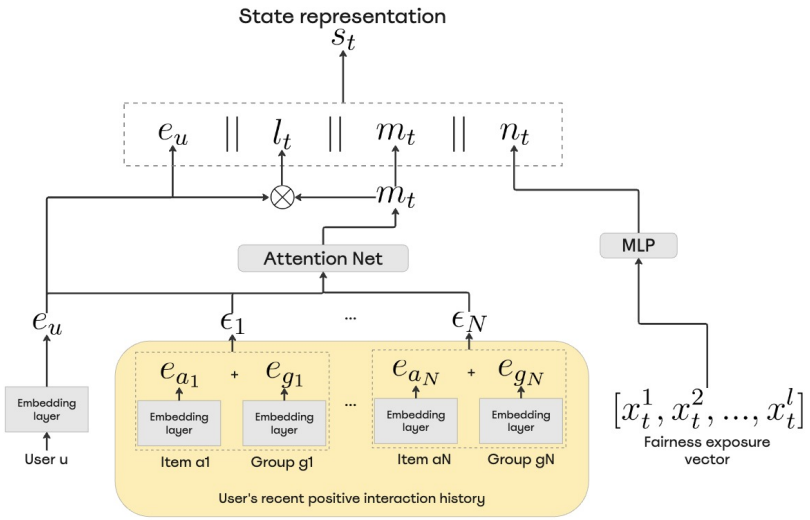

Figure 1: State representation module

State s_t includes the user's recent positive interaction history $H_u = \{a_1, \ldots, a_N\}$, a user representation vector e_u, and current fairness state of the system. A state representation module processes this information into a continuous vector. This module takes as input the last N positively engaged items, and their group identifiers. Each item a is represented by an embedding e_a, and each group g is represented by an embedding e_g (mean of its item embeddings). The item representation $\epsilon = e_a + e_g$ enriches item features with group context (Liu et al., 2020c).

User and item embeddings are pre-trained and fixed during the RL training phase. this has limitations for industrial use, where systems must adapt to changing characteristics, requiring constant retraining, which can be impractical in production (Liu et al., 2020a). However, for the purpose of offline evaluation and demonstrating the core mechanics of A2Fair, fixed pre-trained embeddings suffice.

A2Fair (Figure 2) employs the DDPG Actor-Critic algorithm (Lillicrap et al., 2016). The Actor network is responsible for generating an action a_t given the current state s_t. This state s_t is processed through fully-connected layers to produce an action vector $z_t \in \mathbb{R}^{1\times k}$. This vector z_t then defines a ranking function; the score $score_t$ for a candidate item i_t (represented by its embedding) is calculated as $i_t \cdot z_t^{\intercal}$.. The item with the highest score is then recommended. In this study, one item is recommended at each step. The Critic network evaluates the actions selected by the Actor. It learns to predict the expected long-term return, $Q(s_t, z_t)$, based on the current state s_t and the action vector z_t generated by the Actor (Figure 2).

The training process for A2Fair proceeds over discrete time steps t, with each step involving two primary phases: (1) Transition Generation: The agent first observes the current state s_t. An action a_t, representing the selection of an item, is derived from the Actor network's current policy, with OU-Noise (Uhlenbeck & Ornstein, 1930) for exploration. Following, the reward r_t is estimated using our adaptive reward function (Eq. 5). Subsequently, the user's state (which includes updating the interaction history H_t to H_{t+1} if a positive interaction occurs) and the group exposure metrics are revised, leading to the next state s_{t+1}. The resultant transition tuple (s_t, a_t, r_t, s_{t+1}) is then stored in a replay buffer D. (2) Model Update: From the replay buffer D, a mini-batch of M transitions is sampled, using prioritized experience replay. To further enhance learning stability, the target networks for both the Actor and Critic undergo soft updates.

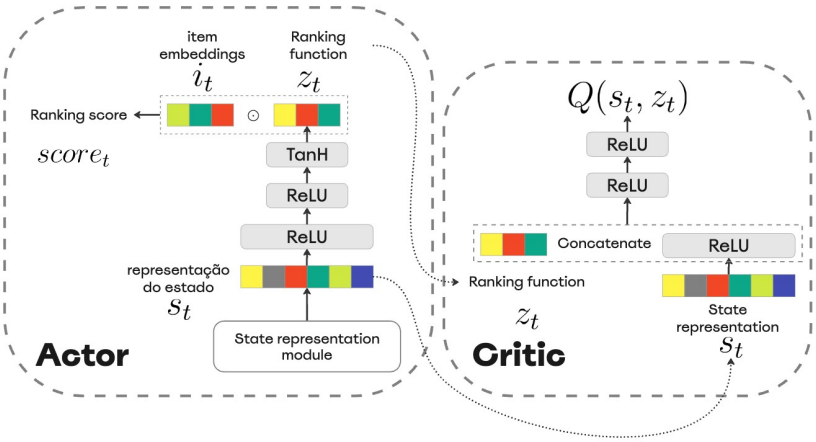

Figure 2: A2Fair Framework

The aforementioned reward r_t is derived from user feedback y_{a_t}. In our offline experimental setup, this feedback is obtained either from historical interactions present within the dataset or, for item-user pairs without existing ratings, simulated. All rewards are subsequently normalized to a consistent range of $[-1, 1]$.

5 Experimental Evaluation

To assess the efficacy of A2Fair, we conducted offline experiments using two public datasets: MovieLens[1] and Yahoo! Music[2]. User ratings, on a 1-5 scale with positive interactions if ≥ 4, are normalized to the range $[-1, 1]$ using the transformation $\frac{1}{2}(\text{rating}_{u,a_t} - 3)$.

- **MovieLens**: Collected by GroupLens Research, with versions from 100k to 1M ratings from the MovieLens platform, supporting RS research.
- **Yahoo! Music (R3)**: Part of Yahoo! Music User Ratings, for large-scale RS evaluation, particularly in music recommendation and user behavior.

We compared A2Fair with conventional, deep learning, RL, and MAB-based algorithms:

- **Random**: Non-personalized baseline, suggests items randomly.
- **MostPopular** (Cañamares & Castells, 2018): Recommends globally most popular items.
- **LinUCB** (Li et al., 2010): Contextual bandit balancing exploration/exploitation.
- **DRR** (Liu et al., 2020b): Deep RL framework for RS, maximizing long-term rewards.
- **FairRec** (Liu et al., 2020c): RL framework balancing accuracy and fairness.

Recommendation effectiveness is measured by $Precision@T$ (fraction of positive interactions among T recommendations), fairness by $PropFair$ (Eq. 2). The trade-off is assessed using the $UFG = \frac{PropFair}{1-Precision}$ score (Liu et al., 2020c), where higher values indicates a better trade-off.

Users were randomly allocated to training (80%) and testing (20%) sets. User and item embeddings (100 dimensions) were pre-trained using PMF (Liu et al., 2020b;c). For simulating user feedback within our offline environment BPMF was selected due to its superior predictive accuracy (Table 1), which is likely attributable to its enhanced modeling of uncertainties.

Table 1: Evaluation of Matrix Factorization methods for simulation.

	PMF	BPMF
	RMSE	
MovieLens (100k)	0.493	**0.477**
MovieLens (1M)	0.464	**0.448**
Yahoo! Music (R3)	0.639	**0.623**

Key hyperparameters for the A2Fair model and training were set as follows: a history length $N = 5$, a discount factor $\gamma = 0.9$, and an episode length $T = 15$. The Actor and Critic networks were constructed with 256-unit hidden layers and optimized using the Adam optimizer with a batch size of 64. To simulate varying levels of item popularity and assess the framework's fairness capabilities under such conditions, items were randomly distributed into five groups of differing sizes. This distribution followed a geometric pattern (Liu et al., 2020c), resulting in an unequal allocation of items. Consequently, some groups inherently contained more items, affording them higher natural visibility, while other groups, with fewer items, faced reduced initial exposure opportunities.

This setup simulates impact of unequal distribution on fairness. Standardizing weights $w_i = 1.0$ for all groups establishes equal importance, allowing analysis of algorithm's ability to promote equity from an unbalanced start. This methodology evaluates the system's efficiency in ensuring equitable group visibility, despite the initial imbalanced distribution.

5.1 Results and Analyses

Table 3 summarizes the simulated online evaluation results, with the best performances for $Precision$ (P), $PropFair$ (F), and UFG (U) metrics highlighted in bold. All results are averaged over 6 independent runs. The LinUCB and DRR baselines perform strongly in terms of recommendation relevance (Precision), which is expected as their primary optimization objective is long-term

[1] https://grouplens.org/datasets/movielens
[2] https://webscope.sandbox.yahoo.com/

Table 3: Experimental results on MovieLens and Yahoo! Music.

	Random	Most Popular	LinUCB	DRR	FairRec	A2Fair
			ML (100k)			
P	$0.775_{\pm 0.05}$	$0.896_{\pm 0.00}$	$0.802_{\pm 0.02}$	$0.908_{\pm 0.02}$	$0.766_{\pm 0.06}$	$\mathbf{0.931}_{\pm 0.01}$
F	$0.871_{\pm 0.00}$	$0.870_{\pm 0.01}$	$0.841_{\pm 0.03}$	$0.860_{\pm 0.01}$	$0.875_{\pm 0.00}$	$\mathbf{0.880}_{\pm 0.00}$
U	$3.866_{\pm 0.6}$	$8.385_{\pm 0.11}$	$4.251_{\pm 0.74}$	$9.923_{\pm 2.45}$	$3.711_{\pm 2.99}$	$\mathbf{12.721}_{\pm 0.42}$
			ML (1M)			
P	$0.725_{\pm 0.01}$	$0.767_{\pm 0.01}$	$0.796_{\pm 0.03}$	$0.784_{\pm 0.02}$	$0.819_{\pm 0.02}$	$\mathbf{0.891}_{\pm 0.03}$
F	$0.872_{\pm 0.01}$	$0.863_{\pm 0.02}$	$0.874_{\pm 0.01}$	$0.876_{\pm 0.01}$	$\mathbf{0.887}_{\pm 0.02}$	$\mathbf{0.887}_{\pm 0.02}$
U	$3.183_{\pm 0.09}$	$0.698_{\pm 0.10}$	$4.276_{\pm 1.92}$	$4.048_{\pm 0.64}$	$4.888_{\pm 0.28}$	$\mathbf{8.138}_{\pm 1.64}$
			YM (R3)			
P	$0.682_{\pm 0.00}$	$0.885_{\pm 0.02}$	$0.810_{\pm 0.02}$	$\mathbf{0.919}_{\pm 0.01}$	$0.903_{\pm 0.01}$	$0.913_{\pm 0.02}$
F	$0.871_{\pm 0.00}$	$0.878_{\pm 0.06}$	$0.841_{\pm 0.02}$	$0.864_{\pm 0.01}$	$\mathbf{0.890}_{\pm 0.00}$	$\mathbf{0.890}_{\pm 0.01}$
U	$2.738_{\pm 0.02}$	$7.639_{\pm 0.20}$	$4.430_{\pm 0.02}$	$\mathbf{10.670}_{\pm 2.84}$	$9.167_{\pm 1.53}$	$10.243_{\pm 5.54}$

reward maximization. In contrast, FairRec demonstrates superior performance in promoting fairness ($PropFair$), though this often comes at the cost of some accuracy.

On both the ML-100k and ML-1M datasets, A2Fair demonstrated superior performance over the baselines across all three evaluated metrics. These findings suggest significant progress in achieving high recommendation relevance without unduly sacrificing fairness, a key practical goal. For the YM(R3) dataset, while DRR achieved the highest raw relevance, it also exhibited the lowest fairness index among the more advanced methods. Conversely, A2Fair achieved a level of fairness comparable to FairRec while simultaneously increasing relevance by 1.1%. On the MovieLens (1M) dataset, A2Fair maintained fairness levels similar to FairRec while achieving an 8.8% gain in relevance. Most notably, on the MovieLens (100k) dataset, A2Fair increased relevance by a substantial 21.5% and fairness by 0.6% compared to FairRec. These results demonstrate A2Fair's capability to efficiently and adaptively balance the often competing objectives of accuracy and fairness.

Table 2: Impact on Relevance and Fairness metrics for different reward formulations within A2Fair.

Reward	Precision	PropFair	UFG
		ML (100k)	
Relevance	**0.948**	0.864	**16.466**
Fairness	0.818	0.864	4.755
Adaptive	0.931	**0.880**	12.721
		ML (1M)	
Relevance	0.864	0.874	6.426
Fairness	0.840	0.882	5.512
Adaptive	**0.891**	**0.887**	**8.138**
		YM (R3)	
Relevance	0.882	0.878	7.441
Fairness	0.837	**0.898**	5.509
Adaptive	**0.913**	0.890	**10.243**

We conducted an ablation study on reward formulations and state representation effectiveness compared to FairRec's approach. First, we studied three reward strategies: maximizing relevance, promoting fairness (Eq. 3), and our adaptive reward (Eq. 5). Table 2 details results. Strategies focused solely on relevance often boost user satisfaction, while an excessive focus on fairness can reduce it.

The adaptive reward formulation in A2Fair demonstrates its effectiveness in dynamically balancing these dimensions. While it may not always achieve the absolute maximum for any single metric (Precision or PropFair), it consistently maintains a strong balance, achieving fairness levels near the ideal without a significant loss in relevance.

We then analyzed the effectiveness of our proposed state representation by comparing it with a configuration inspired by FairRec's approach, which excludes the user embedding from the state. This

comparison, detailed in Table 4, showed that incorporating user features into the state configuration leads to notable improvements.

Finally, we compared our adaptive reward with a formulation inspired by FairRec's bipartite strategy. Contrary to approaches that only assign value to immediate conversions, we argue that from a long-term ecosystem perspective, all exposures hold potential value for future engagement or discovery. Our reward model, which reflects this broader view by valuing all exposures, demonstrated a significant increase in both relevance and fairness metrics (Table 4).

Table 4: Comparing percentage gains in relevance and fairness for A2Fair's proposed components versus alternatives inspired by FairRec's approach (state without user embedding, reward valuing only converting exposures).

	State Representation Gain		Reward Formulation Gain	
	Precision	**PropFair**	**Precision**	**PropFair**
ML (100k)	3.1%	2.0%	8.1%	1.7%
ML (1M)	3.5%	1.7%	8.7%	0.1%
YM (R3)	2.7%	1.1%	2.6%	1.7%

6 Conclusion

We introduced A2Fair, a RL framework designed to balance relevance and exposure fairness in recommender systems. The practical effectiveness of A2Fair is driven by two core design choices: (i) an *adaptive reward function* that dynamically weights relevance and fairness based on measurable user-level diversity preferences, allowing the system to meet individual nuances while pursuing global fairness, and (ii) a *rich state representation* that incorporates user features and the current system-wide exposure distribution, providing the RL agent with comprehensive context for decision-making. Our ablation studies confirm that each of these components independently contributes to improvements in both accuracy and equity.

Although the offline results are encouraging, deploying A2Fair in practice demands further validation and engineering. Our next steps are fourfold. (1) Augment the state encoder with context features (e.g. time of day, trending content). (2) Replace static embeddings with online, jointly trained user- and item-representations that adapt to evolving preferences. (3) Stress-test scalability and decision latency to ensure the system can handle commercial-scale traffic. (4) Run live A/B trials to measure user-perceived relevance, diversity, and fairness, providing the final proof of A2Fair's real-world value.

Broader Impact Statement and Practical Challenges

Deploying adaptive RL for fairness in real-world recommender systems, while promising, introduces non-trivial risks and practical challenges that practitioners must consider. First, optimizing for a specific fairness metric can inadvertently hide other forms of fairness harms or encourage a superficial "checkbox" compliance, rather than genuine equity. The choice and continual evaluation of fairness metrics themselves are critical practical tasks. Second, inherent biases within the training data may be learned and even amplified by the RL agent if embeddings inadvertently capture proxies for sensitive attributes. Third, the increasing complexity of deep RL models can obscure the decision-making process, limiting auditability and making it challenging to explain why a particular recommendation was made. Finally, the very adaptiveness of RL means that rapid online learning and policy changes might, in some cases, destabilize long-term equity goals or lead to unpredictable system behavior if not carefully monitored and constrained. Mitigation requires: (i) multi-metric fairness auditing; (ii) conducting thorough bias detection and mitigation analyses on both input data and learned representations; (iii) developing and utilizing interpretable model diagnostics and explanation techniques; and (iv) implementing continuous monitoring and evaluation frameworks to track downstream effects on users, providers, and the overall platform ecosystem.

Acknowledgments

The authors gratefully acknowledge the valuable insights and constructive discussions provided by Professors Thiago F. Naves and Flávio H. T. Vieira.

This work has been partially funded by the project Research and Development of Digital Agents Capable of Planning, Acting, Cooperating and Learning supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

References

Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1):127–158, Mar 2020. ISSN 1573-1391. DOI: 10.1007/s11257-019-09256-1. URL https://doi.org/10.1007/s11257-019-09256-1.

Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, jun 2005. ISSN 1041-4347. DOI: 10.1109/TKDE.2005.99. URL https://doi.org/10.1109/TKDE.2005.99.

Rocío Cañamares and Pablo Castells. Should i follow the crowd? a probabilistic analysis of the effectiveness of popularity in recommender systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 415–424, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. DOI: 10.1145/3209978.3210014. URL https://doi.org/10.1145/3209978.3210014.

Xiaocong Chen, Lina Yao, Julian McAuley, Guanglin Zhou, and Xianzhi Wang. Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowledge-Based Systems*, 264:110335, 2023. ISSN 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2023.110335. URL https://www.sciencedirect.com/science/article/pii/S0950705123000850.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020a.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2185–2195. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/creager20a.html.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 525–534, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. DOI: 10.1145/3351095.3372878. URL https://doi.org/10.1145/3351095.3372878.

Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: Scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pp. 271–280, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936547. DOI: 10.1145/1242572.1242610.

Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, Apr 2023. ISSN 1573-1391. DOI: 10.1007/s11257-023-09364-z. URL https://doi.org/10.1007/s11257-023-09364-z.

Gabriel Dulac-Arnold, Richard Evans, Peter Sunehag, and Ben Coppin. Reinforcement learning in large discrete action spaces. *CoRR*, abs/1512.07679, 2015. URL http://arxiv.org/abs/1512.07679.

Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pp. 445–453, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. DOI: 10.1145/3437963.3441824. URL https://doi.org/10.1145/3437963.3441824.

Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pp. 316–324, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. DOI: 10.1145/3488560.3498487. URL https://doi.org/10.1145/3488560.3498487.

Usman Gohar, Zeyu Tang, Jialu Wang, Kun Zhang, Peter L. Spirtes, Yang Liu, and Lu Cheng. Long-term fairness inquiries and pursuits in machine learning: A survey of notions, methods, and challenges, 2025. URL https://arxiv.org/abs/2406.06736.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/eb163727917cbba1eea208541a643e74-Paper.pdf.

Michal Kompan and Mária Bieliková. Content-based news recommendation. In Francesco Buccafurri and Giovanni Semeraro (eds.), *E-Commerce and Web Technologies*, pp. 61–72, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15208-5.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. DOI: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun (eds.), *ICLR*, 2016. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2016.html#LillicrapHPHETS15.

Feng Liu, Huifeng Guo, Xutao Li, Ruiming Tang, Yunming Ye, and Xiuqiang He. End-to-end deep reinforcement learning based recommendation with supervised embedding. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pp. 384–392, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450368223. DOI: 10.1145/3336191.3371858. URL https://doi.org/10.1145/3336191.3371858.

Feng Liu, Ruiming Tang, Xutao Li, Weinan Zhang, Yunming Ye, Haokun Chen, Huifeng Guo, Yuzhou Zhang, and Xiuqiang He. State representation modeling for deep reinforcement learning based recommendation. *Knowledge-Based Systems*, 205:106170, 2020b. ISSN 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2020.106170. URL https://www.sciencedirect.com/science/article/pii/S095070512030407X.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 6196–6200. International Joint Conferences on Artificial Intelligence Organization, 7 2019. DOI: 10.24963/ijcai.2019/862. URL https://doi.org/10.24963/ijcai.2019/862.

Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. Balancing between accuracy and fairness for interactive recommendation with reinforcement learning. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I*, pp. 155–167, Berlin, Heidelberg, 2020c. Springer-Verlag. ISBN 978-3-030-47425-6. DOI: 10.1007/978-3-030-47426-3_13. URL https://doi.org/10.1007/978-3-030-47426-3_13.

Masoud Mansoury and Bamshad Mobasher. Fairness of exposure in dynamic recommendation, 2023. URL https://arxiv.org/abs/2309.02322.

Masoud Mansoury, Finn Duijvestijn, and Imane Mourabet. Potential factors leading to popularity unfairness in recommender systems: A user-centered analysis. *arXiv preprint arXiv:2310.02961*, 2023.

Benjamin Marlin and Richard S. Zemel. The multiple multiplicative factor model for collaborative filtering. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 73, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. DOI: 10.1145/1015330.1015437.

Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness and satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 2243–2251, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. DOI: 10.1145/3269206.3272027. URL https://doi.org/10.1145/3269206.3272027.

Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S. Thomas. Offline contextual bandits with high probability fairness guarantees. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d69768b3da745b77e82cdbddcc8bac98-Paper.pdf.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Owen Phelan, Kevin Mccarthy, Mike Bennett, and Barry Smyth. Terms of a feather: Content-based news recommendation and discovery using twitter. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval - Volume 6611*, ECIR 2011, pp. 448–459, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642201608. DOI: 10.1007/978-3-642-20161-5_44.

Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommendations: An overview. *The VLDB Journal*, 31(3):431–458, oct 2021. ISSN 1066-8888. DOI: 10.1007/s00778-021-00697-y. URL https://doi.org/10.1007/s00778-021-00697-y.

Juno Prent and Masoud Mansoury. Correcting popularity bias in recommender systems via item loss equalization, 2025. URL https://arxiv.org/abs/2410.04830.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pp. 2219–2228, New York, NY, USA, 2018. Association for Computing Machinery. ISBN

9781450355520. DOI: 10.1145/3219819.3220088. URL https://doi.org/10.1145/3219819.3220088.

Ashudeep Singh and Thorsten Joachims. *Policy Learning for Fairness in Ranking*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009, jan 2009. ISSN 1687-7470. DOI: 10.1155/2009/421425. URL https://doi.org/10.1155/2009/421425.

Haoran Tang, Shiqing Wu, Zhihong Cui, Yicong Li, Guandong Xu, and Qing Li. Model-agnostic dual-side online fairness learning for dynamic recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2727–2742, 2025. DOI: 10.1109/TKDE.2025.3544510.

G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Phys. Rev.*, 36:823–841, Sep 1930. DOI: 10.1103/PhysRev.36.823. URL https://link.aps.org/doi/10.1103/PhysRev.36.823.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10686–10696. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21b.html.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, 41(3), feb 2023. ISSN 1046-8188. DOI: 10.1145/3547333. URL https://doi.org/10.1145/3547333.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1144–1152. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/wen21a.html.

Himank Yadav, Zhengxiao Du, and Thorsten Joachims. Policy-gradient training of fair and unbiased ranking functions. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 1044–1053, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. DOI: 10.1145/3404835.3462953. URL https://doi.org/10.1145/3404835.3462953.

Xiwang Yang, Yang Guo, Yong Liu, and Harald Steck. A survey of collaborative filtering based social recommender systems. *Computer Communications*, 41:1–10, 2014. ISSN 0140-3664. DOI: https://doi.org/10.1016/j.comcom.2013.06.009. URL https://www.sciencedirect.com/science/article/pii/S0140366413001722.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellström, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pp. 1040–1048, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. DOI: 10.1145/3219819.3219886. URL https://doi.org/10.1145/3219819.3219886.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. Drn: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pp. 167–176, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. DOI: 10.1145/3178876.3185994. URL https://doi.org/10.1145/3178876.3185994.

Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pp. 2810–2818, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. DOI: 10.1145/3292500.3330668. URL https://doi.org/10.1145/3292500.3330668.