1. Introduction

Achieving fair and engaging gameplay across heterogeneous player classes remains a persistent challenge in collectible card games (CCGs). In Hearthstone, the presence of 11 distinct classes with highly asymmetric card pools ranging from NEUTRAL (758 cards) to DREAM (2 cards), frequently results in persistent win-rate imbalances. Traditional manual balancing by designers is time-intensive and difficult to scale with frequent card releases. This work proposes a **MULTI-AGENT REINFORCEMENT LEARNING (MARL)** framework in which each Hearthstone class is represented by an independent agent that learns to modify its own card parameters (attack, health, mana cost) to optimise a global fairness objective. The approach draws upon five key contributions from the MARL and imperfect-information games literature.

2. Related Work

2.1 Simulation Environments for Imperfect-Information Card Games

Zha et al. (2019) introduced **RLCard**[1], an open-source toolkit that provides standardised OpenAI Gym environments for a variety of imperfect-information card games, including Texas Hold'em, Dou Dizhu, UNO, and Mahjong. The toolkit addresses the combinatorial explosion of states (up to 10^{163} in UNO) and actions through configurable state abstractions and action masking, while supporting native multi-agent self-play. RLCard has become a de-facto platform for benchmarking MARL algorithms in card-game settings due to its modular design and comprehensive documentation.

The present work extends the Dou Dizhu environment within RLCard to create HearthstoneBalanceEnv, preserving the original toolkit's handling of hidden information while introducing class-specific action spaces and a belief-state representation derived from Monte Carlo sampling of the unobserved card pool.

2.2 Competitive Adaptation in Multi-Player Card Games

Barros et al. (2021) conducted a systematic empirical comparison of three major reinforcement learning algorithms in Chef's Hat, a four-player competitive card game originally developed for human and robot interaction research. The algorithms examined were **Deep Q-Learning (DQL)**, a value-based off-policy method that learns an action value function via a deep neural network and experience replay buffer; **Advantage Actor-Critic (A2C)**, an on-policy actor critic approach that trains a policy network (the actor) alongside a state value estimator (the critic) to produce low-variance policy gradient updates; and **Proximal Policy Optimization (PPO)**, a trust-region policy gradient technique that enforces stable learning by clipping the probability ratio between old and new policies, achieving considerably better sample efficiency and robustness. Although all three algorithms eventually converged to near uniform win rates of approximately 25 % per player under self-play, PPO demonstrated substantially faster adaptation to shifting opponent strategies and markedly lower variance throughout training.

---

[1] https://github.com/datamllab/rlcard

These findings provide strong justification for adopting **Proximal Policy Optimization (PPO)** as the core learning algorithm for the eleven class-specific agents in the present Hearthstone balancing framework. To reorient optimisation from individual victory toward collective fairness, the binary win/loss reward originally used by Barros et al. (2021) is replaced with Jain's Fairness Index calculated across all class win rates, augmented by a modest corrective term that prevents any class from becoming permanently suppressed.

2.3 Partially Observable Markov Decision Processes in Card Games

One of the earliest applications of reinforcement learning to multi-player imperfect-information card games is presented by Fujita et al. (2003), who formulated the game of Hearts as a **Partially Observable Markov Decision Process (POMDP)**. By estimating hidden card distributions and employing mean-field opponent modelling, their agent outperformed strong rule-based opponents after 80 000 training episodes.

The POMDP framework is adopted to represent Hearthstone's hidden decks. Monte Carlo sampling (100 iterations) from the remaining card pool generates an approximate belief over opponent class composition, which is subsequently refined by a lightweight neural density estimator trained on historical game trajectories. This constitutes a substantial improvement over the mean-field approximation used by Fujita et al. (2003), providing more accurate probabilistic observations for downstream policy learning.

2.4 Bio-Inspired and Evolutionary MARL

Li et al. (2025a) provide a comprehensive review that integrates MARL with game-theoretic foundations and bio-inspired computation, establishing a methodological taxonomy spanning value-function decomposition, policy gradients, and online search planning. Population-based training (PBT) and evolutionary strategy generation are highlighted as particularly effective for exploration in high-dimensional, non-stationary environments.

These principles are operationalised through a population of eight PPO variants per class. Every 5000 training episodes, agents are evaluated according to Jain's Fairness Index; the top two performers are retained while the remainder undergo hyperparameter mutation and crossover. A shared central critic exposes the complete vector of class win rates during training (centralised training with decentralised execution, CTDE), enabling coordinated adjustments toward global equilibrium.

2.5 MARL in Modern Video Games and Fairness Evaluation

Li et al. (2025b) offer a broad survey of MARL applications from AlphaStar (DeepMind, 2019) and OpenAI Five (OpenAI, 2019) to contemporary Multiplayer Online Battle Arena (MOBA) and Real-Time Strategy (RTS) titles. Persistent challenges — non-stationarity, partial observability, sparse rewards, and coordination — are identified, alongside a quantitative game-complexity metric and an explicit call for standardised fairness evaluation in multiplayer settings.

Training follows the CTDE paradigm to mitigate non-stationarity. A Hearthstone-specific complexity score

1. $C = \log_{10}(11) \times$ (average cards per class) $\times$ (hidden-information factor)

is incorporated as a reward penalty. Post-training evaluation on 10 000 independent matches reports Jain's Fairness Index, win-rate variance, and standard deviation (target < 0.03), directly addressing the standardised fairness metrics advocated by Li et al. (2025b).

Summary

Taken together, the five studies form a clear and coherent lineage that directly enables the present work, yet none of them addresses the specific challenge of automated, fairness-driven balance in a large-scale asymmetric collectible card game.

RLCard (Zha et al., 2019) provides the essential simulation infrastructure, but its environments are designed for fixed rules and individual performance rather than for agents that actively reshape the game's parameters.

Barros et al. (2021) convincingly show that PPO is the most reliable algorithm when multiple agents must continually adapt to one another's strategies, yet their reward remains a zero-sum win signal instead of an explicit equity objective.

Fujita et al. (2003) demonstrate that POMDP-style belief modelling is both necessary and feasible in multi-player card games with hidden information, but the scale and asymmetry of modern CCGs far exceed the four-player Hearts.

The recent reviews by Li et al. (2025a, 2025b) underline the maturity of population-based training, centralised training with decentralised execution, and the urgent need for rigorous fairness metrics in multiplayer settings—yet they stop short of applying these ideas to the problem of class balance itself.

The contribution of this project therefore lies in the deliberate synthesis of these strands: eleven independent PPO agents, each responsible for one Hearthstone class, are placed in a shared RLCard-based environment where the sole scalar reward is Jain's Fairness Index over the global vector of class win rates. Hidden deck composition is modelled with Monte-Carlo-sampled beliefs refined by a learned density estimator, population-based training preserves behavioural diversity across generations, and a central critic provides the coordination signal required for stable convergence. The result is a principled, reproducible pipeline that moves the balance process from subjective designer judgement to an optimisable, data-driven objective—directly closing the gap identified across the surveyed literature between general MARL capability and the practical requirements of live-service collectible card games.

References

1. Zha, D., Lai, K.H., Cao, Y., Huang, S., Wei, R., Guo, J. and Hu, X., 2019. Rlcard: A toolkit for reinforcement learning in card games. arXiv preprint arXiv:1910.04376.
2. Barros, P., Tanevska, A. and Sciutti, A., 2021, January. Learning from learners: Adapting reinforcement learning agents to be competitive in a card game. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 2716-2723). IEEE.

3. Fujita, H., Matsuno, Y. and Ishii, S., 2003, October. A reinforcement learning scheme for a multi-agent card game. In SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483) (Vol. 5, pp. 4071-4078). IEEE.
4. Li, H., Yang, P., Liu, W., Yan, S., Zhang, X. and Zhu, D., 2025. Multi-Agent Reinforcement Learning in Games: Research and Applications. Biomimetics, 10(6), p.375.
5. Li, Z., Ji, Q., Ling, X. and Liu, Q., 2025. A comprehensive review of multi-agent reinforcement learning in video games. IEEE Transactions on Games.