# A Comparative Analysis of Reinforcement Learning Training Strategies for Agents on Competitive 2v2 Soccer

**Marcelo P. Rezende**[1], **Rafael P. Torchelsen**[1,2]

[1]Centro de Desenvolvimento Tecnológico - Universidade Federal de Pelotas (UFPel)

[2]AKCIT Lab - Instituto de Informática - UFRGS

{marcelo.pr,rafael.torchelsen}@inf.ufpel.edu.br

***Abstract. Introduction:*** *The use of Reinforcement Learning (RL) in digital games has grown significantly, enabling agents to learn complex behaviors through interaction, with many approaches emphasizing improvements in agent performance. However, in game development, aspects such as perceived difficulty, behavioral naturalness, and player enjoyment are equally important.* ***Objective:*** *This study aimed to analyze the impact of different RL training strategies on both objective performance and subjective player experience in a competitive game scenario.* ***Methodology or Steps:*** *A 2v2 soccer environment was developed in Unity using the ML-Agents toolkit. The original scene was modified to include deterministic agents with scalable difficulty. Three models were trained using the multi-agent Posthumous Credit Assignment (POCA) algorithm: one with self-play and two against deterministic opponents with difficulty adjustment based on win streak and win rate. For evaluation, pairs of human players participated in sessions of three matches, each against a different model, followed by a questionnaire assessing perceived difficulty, movement naturalness, and enjoyment. Objective gameplay metrics, including match results and movement heatmaps, were also collected.* ***Results:*** *The self-play model was consistently perceived as the most difficult. In contrast, models trained against deterministic opponents with difficulty scaling were rated as more natural and more enjoyable. These results indicate that higher competitive performance does not necessarily lead to better player experience. Limitations include the small sample size and evaluation in a single scenario. Future work involves testing additional RL algorithms and extending the environment to more complex or varied game contexts.*

***Keywords*** *Reinforcement Learning, self-play, Curriculum Learning, Video Games, User Experience.*

## 1. Introduction

The use of Artificial Intelligence (AI) techniques in digital games has become increasingly important, enabling the development of more dynamic and adaptive Non-Playable Characters (NPCs). Traditional approaches, such as Finite State Machines (FSMs) and Behavior Trees (BTs), provide developers with full control over agent behavior, but often result in predictable and repetitive patterns [Yannakakis e Togelius 2025].

Recent advances in Reinforcement Learning (RL) [Sutton e Barto 2018] have introduced new possibilities, allowing agents to learn complex behaviors through interaction with the environment [Haddad et al. 2021]. In particular, multi-agent

Reinforcement Learning (MARL) has shown strong potential in competitive and cooperative scenarios, where multiple agents must learn to coordinate or compete simultaneously [Albrecht et al. 2024] [Huh e Mohapatra 2024]. Despite these advances, most research in RL for games still focuses primarily on objective performance metrics, such as win rate and score.

However, in the context of digital games, performance alone is not sufficient. Player experience factors, such as perceived difficulty, behavioral naturalness, and enjoyment, play a crucial role in determining the quality of AI-driven gameplay. Highly optimized agents may lead to unbalanced or frustrating experiences, highlighting a gap between technical performance and player satisfaction.

In this context, we investigate how different MARL training strategies affect not only objective performance but also the subjective player experience. A 2v2 soccer environment was developed using Unity ML-Agents [1], where three models were trained using the multi-agent Posthumous Credit Assignment (POCA) [Cohen et al. 2022] algorithm under different training conditions: self-play [DiGiovanni e Zell 2021] and training against deterministic opponents with adaptive difficulty mechanisms.

The models were evaluated through an experimental study with human participants, combining objective gameplay metrics with subjective assessments of difficulty, naturalness, and enjoyment. Our results show that agents optimized purely for competitive performance are not necessarily perceived as the most enjoyable to play against, suggesting important trade-offs for the design of game AI.

The main contributions of this work are a comparative analysis of MARL training strategies in a controlled game environment, an evaluation methodology combining objective and subjective metrics, along with empirical evidence highlighting the disconnect between performance optimization and player experience.

## 2. Related Works

Recent advances in RL and MARL have enabled the development of increasingly complex agents for digital games. Prior work in this area can be roughly organized into three main directions: MARL applications in games, training strategies such as self-play and curriculum learning, and evaluation of agent behavior from a human-centered perspective.

In the context of MARL for games, [Li et al. 2025] provides a comprehensive survey highlighting key challenges such as non-stationarity and credit assignment, as well as successful applications in competitive environments like sports simulations. Similarly, [Huynh et al. 2025] proposes a hybrid training approach combining curriculum learning and population-based self-play in a 2v2 Pommerman environment, demonstrating that adaptive difficulty and matchmaking strategies can significantly improve agent performance.

Regarding training strategies, [Baker et al. 2020] shows that self-play can generate emergent behaviors through an automatic curriculum driven by competition, leading to increasingly complex strategies without explicit supervision. In contrast, [Almeida et al. 2024] demonstrates that curriculum learning can outperform imitation-

[1] https://github.com/unity-technologies/ml-agents

based approaches in competitive first-person shooter scenarios, reinforcing the importance of structured training progression.

From a human-centered evaluation perspective, [Milani et al. 2023] investigates how players perceive human-likeness in agent behavior, showing that movement smoothness and goal-directed actions are key factors in realism. Likewise, [Servat e Mohamadi 2023] demonstrates that RL-trained agents can exhibit more natural and less predictable behaviors than rule-based systems, thereby improving immersion.

Despite these advances, most studies still focus either on objective performance metrics or on behavioral realism, often treating them separately. Few studies explicitly investigate the relationship between training strategies, competitive performance, and subjective player experience in multi-agent game scenarios. This gap motivates the present work, which evaluates different MARL training approaches by jointly considering objective metrics and human perception.

## 3. Test Bed

The experiments were conducted in a 2v2 soccer environment based on the SoccerTwos scenario provided by the Unity ML-Agents toolkit (Release 23) [2] (Figure 1). This physics-based environment represents a cooperative-competitive multi-agent setting in which two teams interact in real time to score goals. The scenario was selected due to its balance between complexity and controllability, allowing the emergence of coordination, positioning, and adversarial behaviors while maintaining a manageable state and action space. [3]

To accelerate training, eight parallel instances of the environment were executed simultaneously, enabling efficient experience collection.

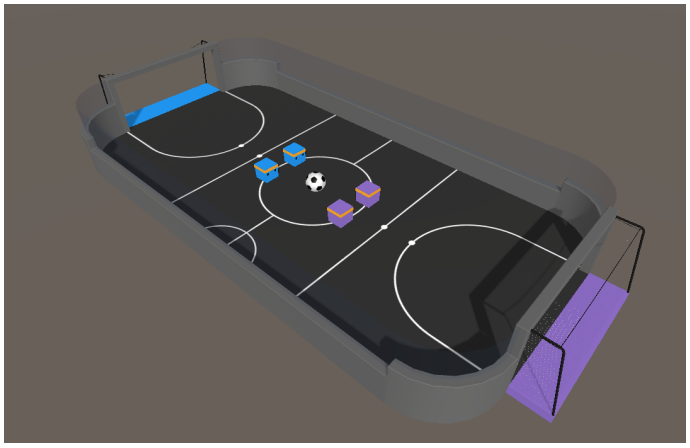

**Figure 1. Soccer simulation environment provided by Unity ML-Agents**

Each agent receives a 336-dimensional vector observation generated through a ray-casting perception system composed of 14 rays: 11 forward-facing rays distributed

[2]The projects used for model training and gameplay during experiments is available at: `https://github.com/MarceloFaiz/ML-Agents-Modified-SoccerTwos`

[3]A video demonstration of the environment and experimental setup is available at: `https://osf.io/r6awv/overview?view_only=6aaf69b33f824777b8f4c537fdc49049`

over a 120° field of view and 3 backward-facing rays over 90° (Figure 2). Each ray encodes the type of detected object (e.g., ball, teammate, opponent, wall, or goal) and its normalized distance. This representation provides a computationally efficient alternative to visual input while preserving spatial awareness.

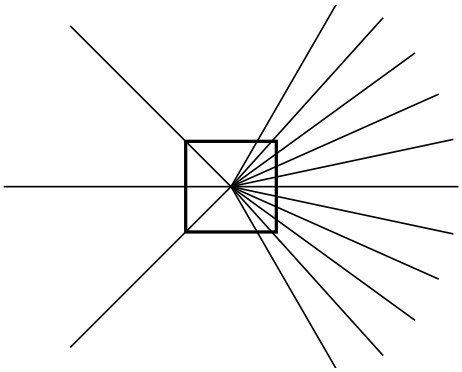

**Figure 2. Top-down representation of the agent's ray-cast observation model.**

The action space is defined as a discrete multi-branch structure, allowing agents to control movement and rotation simultaneously. This enables reactive and continuous navigation while keeping the action representation tractable.

The reward function was designed to balance task completion with temporal efficiency, combining sparse and shaped signals:

- $+(1 - t_{penalty})$ when a goal is scored;
- $-1$ when a goal is conceded;

where $t_{penalty} = t/T$ and $T = 25.000$ environment steps per episode, such that $t_{penalty} \in [0, 1]$. The time penalty increases over the course of an episode, decreasing the reward for scoring as the episode progresses and encouraging faster goal completion.

To support controlled training conditions, this work adapted the default heuristic controller provided by the environment scripts with a set of manually designed deterministic policies. While the original heuristic implementation allows human control via keyboard input, it was extended in this work to implement rule-based agents with increasing levels of difficulty.

These policies are not learned; instead, they are explicitly programmed to exhibit progressively more complex behaviors. This design enables precise control over opponent difficulty and allows the construction of structured training curriculum:

- **Inactive:** NPCs remain stationary, providing a minimal baseline for initial learning.
- **Easy:** NPCs follow a simple reactive policy, moving directly toward the ball without strategic positioning.
- **Medium:** NPCs pursue the ball while orienting their movement toward the opponent's goal, introducing a basic notion of offensive direction.
- **Hard:** NPCs adopt a fixed role-based strategy, where one agent acts as an attacker (ball pursuit and shooting) and the other as a defender (goal protection and ball interception).

Such variations enable the implementation of adaptive training regimes in which opponent difficulty is dynamically adjusted based on the agent's performance (Figure 3). Unlike self-play, where opponent complexity emerges implicitly over time, this approach provides explicit control over the learning progression, allowing the difficulty to be increased or decreased according to predefined criteria.

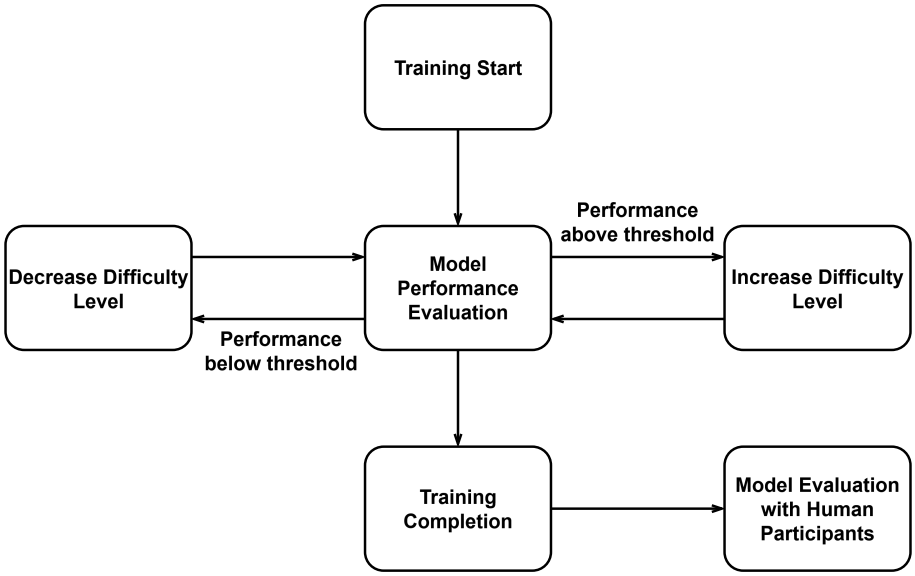


**Figure 3. Training against deterministic opponents**

### 3.1. Training Strategies

Three distinct strategies were employed:

- **self-play:** Trained by competing against copies of itself. It uses a buffer of previous checkpoints to maintain opponent diversity and stabilize learning.
- **WinStreak:** Trained against the ladder of algorithms. The difficulty increases after 8 consecutive wins and decreases after 5 consecutive losses.
- **WinRate:** Trained against the ladder of algorithms. The difficulty increases when the win rate, over 100 matches, exceeds 70% and decreases when it falls below 40%.

The thresholds used in the WinStreak and WinRate strategies were empirically defined based on preliminary exploratory tests. Very low values made the difficulty progression more unstable, while very high values significantly hindered transitions between levels. After experimentation, the adopted values ensured stable progression without excessive oscillations between difficulty levels.

### 3.2. Learning Algorithms and Training Setup

Proximal Policy Optimization (PPO) [Schulman et al. 2017] is a widely used policy-gradient algorithm that serves as the foundation for several modern reinforcement learning methods. It follows an actor-critic paradigm, where a policy $\pi_\theta(a|s)$ selects actions and a value function $V(s)$ estimates expected returns.

A key feature of PPO is the use of a clipped surrogate objective, which constrains policy updates to remain close to the previous policy. This mechanism

prevents excessively large parameter updates, improving training stability and avoiding performance collapse during learning. As a result, PPO achieves a practical balance between sample efficiency, robustness, and ease of implementation, making it a standard choice for complex environments such as games and multi-agent systems.

To address the challenges of multi-agent reinforcement learning, this work employs the multi-agent Posthumous Credit Assignment algorithm, an extension of PPO designed for cooperative settings [Cohen et al. 2022].

POCA introduces a centralized critic during training, which has access to the joint observations and actions of all agents. This design enables more accurate credit assignment by estimating each agent's contribution to the shared reward, a key difficulty in multi-agent environments. At execution time, however, policies remain decentralized, allowing each agent to act independently based on local observations.

By combining centralized training with decentralized execution, POCA improves coordination and learning efficiency, enabling the emergence of cooperative behaviors in dynamic and partially observable environments [Cohen et al. 2022].

All models were trained using the official POCA trainer implementation provided by the Unity ML-Agents Toolkit (Release 23), using the default training configuration supplied for the SoccerTwos environment. For models trained against deterministic opponents, only the self-play component was disabled. All models shared the same network architecture and hyperparameters to ensure a fair comparison (Table 1). Although the maximum number of steps was set to 50 million, training was manually stopped at approximately 20 million steps based on observed convergence.

**Table 1. Training hyperparameters (POCA)**

| Parameter | Value |
|---|---|
| *Batch size* | 2048 |
| *Buffer size* | 20480 |
| *Learning rate* | $3 \times 10^{-4}$ |
| *Entropy coefficient* ($\beta$) | 0.005 |
| *Clipping parameter* ($\epsilon$) | 0.2 |
| *GAE* ($\lambda$) | 0.95 |
| *Discount factor* ($\gamma$) | 0.99 |
| *Num epochs* | 3 |
| *Hidden units* | 512 |
| *Num layers* | 2 |
| *Time horizon* | 1000 |
| *Max steps* | $5 \times 10^{7}$ |

All experiments were conducted on a machine equipped with an AMD Ryzen 5 3600 CPU, an NVIDIA GeForce RTX 2060 GPU, and 16 GB of DDR4 RAM. Under this setup, the training process took approximately 12 hours on average for each model.

### 3.3. Experiment

An experimental protocol was designed to evaluate both agent performance and player perception under controlled and reproducible conditions. The study combined gameplay

metrics with subjective evaluation, enabling analysis of performance as well as perceived difficulty, naturalness, and enjoyment. All participants were exposed to the same conditions to ensure comparability across models.

To conduct the experiments, the simulation environment was extended with features required for interactive gameplay, including a graphical interface, match timing, score display, and automated data logging. Together, these features allowed the environment to be used as a playable game while preserving its underlying simulation dynamics.

The study involved 15 human pairs facing trained agent teams in the environment's 2v2 setup. Participants were recruited voluntarily in an academic setting without gaming experience restrictions. Prior to participation, all individuals were informed about the study's purpose and procedures. No personally identifiable information was collected, and all recorded data were fully anonymized. Before testing, pairs completed a brief practice match to minimize control-familiarity bias. Because every pair evaluated all models, individual skill variations affected all conditions equally.

Each session comprised three randomized 1-minute matches against each evaluated model once, balancing interaction depth with fatigue prevention. Sessions were assigned unique identifiers, automatically generating log files (model type, outcome, final score, and timestamps) and movement heatmaps to analyze player trajectories and spatial behavior.

Following the matches, participants completed a survey evaluating each model on a 5-point Likert scale across perceived difficulty, agent naturalness, and overall enjoyment. Finally, participants provided a comparative ranking to identify which model was the most difficult, natural, and enjoyable overall.

## 4. Results

This section presents the results obtained from experiments conducted with 15 pairs of participants. The analysis combines objective performance metrics with subjective user evaluations to investigate the impact of different training strategies on both agent effectiveness and player experience.

Objective performance was evaluated based on match outcomes, including goals scored, goals conceded, and win rate against human players (Table 2).

**Table 2. Summary of AI Models Performance**

| Model | Goals Scored | Goals Against | W-D-L | Win Rate |
|---|---|---|---|---|
| *self-play* | 101 | 13 | 14-0-1 | 93.33% |
| *WinStreak* | 54 | 10 | 15-0-0 | 100% |
| *WinRate* | 53 | 31 | 10-1-4 | 66.67% |

The self-play model was clearly the most dominant in terms of performance, achieving 14 wins out of 15 matches and a substantially higher goal difference than the other models. This suggests that the learned policy is highly optimized for competitive play and capable of consistently outperforming human players. The win-streak model also showed strong performance, remaining undefeated across all matches. However,

its offensive output was considerably lower than that of the self-play model, suggesting a more conservative or controlled strategy. In contrast, the win-rate model showed the weakest performance among the three, with a lower win rate and a higher number of goals conceded. Despite remaining competitive, its results indicate reduced consistency and greater defensive vulnerability.

Subjective metrics were collected through participant questionnaires, evaluating perceived difficulty, behavioral naturalness, and enjoyment using a five-point Likert scale (Figure 4), as well as direct comparative choices (Figure 5).

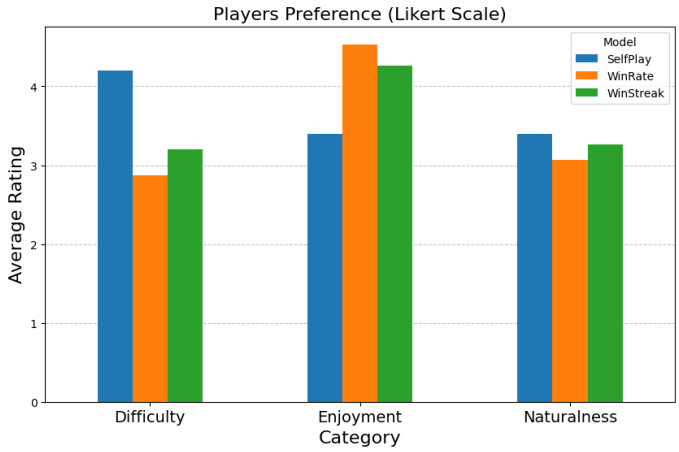


**Figure 4. Average player ratings (Likert scale) for difficulty, enjoyment, and naturalness across the evaluated models**

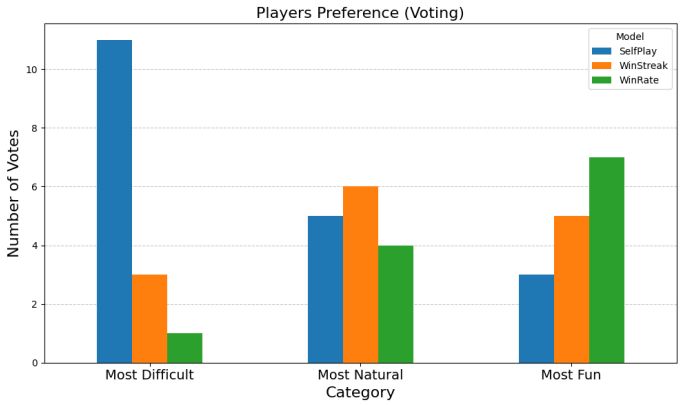


**Figure 5. Player preferences based on voting for the most difficult, most natural, and most fun models**

The self-play model was consistently perceived as the most difficult, both in average ratings and in direct comparisons. This aligns with its superior objective performance. Regarding naturalness, the results were more balanced. While the self-play model achieved slightly higher average ratings, the win-streak model was more frequently selected as the most natural in direct comparisons, suggesting subtle differences in perceived behavior quality. In terms of enjoyment, an inverse trend was observed. The win-rate model was rated as the most enjoyable, followed by the win-streak model, while

the self-play model received the lowest enjoyment scores. This reinforces the idea that excessive difficulty can negatively impact player experience.

The results reveal important trade-offs between competitive performance, perceived naturalness, and player enjoyment. The self-play model achieved the highest level of objective performance, exhibiting a highly aggressive and dominant behavior. Analysis of movement patterns indicates frequent offensive positioning and sustained pressure on the opposing team (Figure 6).

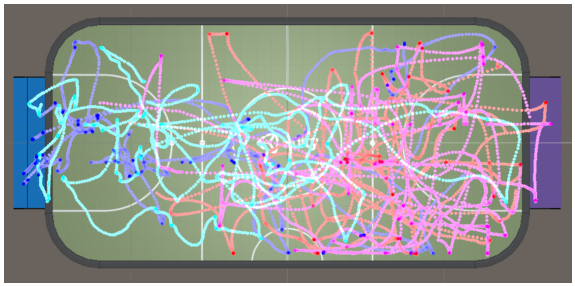

**Figure 6. Heatmap showing movements of self-play team (blue and cyan) and human players (red and pink)**

While this behavior contributed to its high win rate and goal difference, it also resulted in increased perceived difficulty and reduced enjoyment. This suggests that agents optimized purely for performance may lead to unbalanced and potentially frustrating gameplay experiences. In contrast, models trained against heuristic opponents exhibited behaviors more aligned with desirable player experience. The win-streak model maintained strong performance while presenting more controlled gameplay dynamics. The win-rate model, despite lower performance, provided a more balanced and engaging experience.

Overall, these observations suggest that less dominant agents, especially those shaped by adaptive difficulty, tend to produce more enjoyable interactions by allowing players greater participation and perceived agency.

Behavioral differences between models were further analyzed based on match statistics and gameplay patterns, including score margins and observed movement distributions. The self-play model exhibited predominantly proactive behavior, characterized by continuous offensive pressure and large-margin victories (Figure 7). This suggests that the agent actively seeks to maximize reward even after achieving a winning state. In contrast, the win-streak and win-rate models demonstrated more reactive behavior. Their gameplay patterns indicate a tendency to respond to opponent actions and maintain advantage rather than aggressively expanding it. This results in more balanced matches, which may contribute to higher enjoyment ratings.

## 5. Conclusion

In this work, we investigated the impact of different MARL training strategies on both objective performance and player experience in a competitive 2v2 game scenario. Three models were trained using distinct approaches and evaluated through a combination of gameplay metrics and human-centered assessments.

```
SESSION ID: 7DGBAV
DATE:
================================================
Match #1
   Opponent Model: NewStreak8or5-19999917
   Final Score:    Purple 1 - 3 Blue
   Winner:         Blue (CPU)
------------------------------------------------
Match #2
   Opponent Model: NewWinrate7or4-19999978
   Final Score:    Purple 3 - 2 Blue
   Winner:         Purple (Player)
------------------------------------------------
Match #3
   Opponent Model: SelfPlay-19999972
   Final Score:    Purple 1 - 7 Blue
   Winner:         Blue (CPU)
------------------------------------------------
SESSION END:
```

**Figure 7. Text log generated after a session.**

Our results show that higher competitive performance does not necessarily translate into a better player experience. The self-play model achieved the strongest objective results, consistently outperforming human players and being perceived as the most difficult opponent. However, this dominance was associated with lower enjoyment scores, indicating that excessive difficulty may negatively affect player engagement.

In contrast, models trained against adaptive opponents produced more balanced interactions. Although less dominant in terms of performance, these models were perceived as more enjoyable and, in some cases, more natural. These findings highlight the importance of explicitly considering player experience as a central factor in the design of game AI, rather than focusing solely on maximizing performance metrics.

The main contribution of this work lies in the combined evaluation of MARL training strategies using both objective and subjective metrics, providing empirical evidence of the trade-offs between competitiveness and player satisfaction in multi-agent game environments.

Future work includes evaluating additional reinforcement learning algorithms, exploring real-time adaptive behaviors, and extending the experiments to more complex environments and different game genres. Incorporating human-in-the-loop training strategies may also further improve the naturalness and perceived quality of agent behavior.

## 6. Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and partially by CNPq–Brazil (grant 421962/2023-2), and by the 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

Albrecht, S. V., Christianos, F., e Schäfer, L. (2024). *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, Cambridge, MA, USA.

Almeida, P., Carvalho, V., e Simões, A. (2024). Reinforcement learning as an approach to train multiplayer first-person shooter game agents. *Technologies*, 12(3):34.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., e Mordatch, I. (2020). Emergent tool use from multi-agent autocurricula. arXiv:1909.07528.

Cohen, A., Teng, E., Berges, V.-P., Dong, R.-P., Henry, H., Mattar, M., Zook, A., e Ganguly, S. (2022). On the use and misuse of absorbing states in multi-agent reinforcement learning. arXiv:2111.05992.

DiGiovanni, A. e Zell, E. C. (2021). Survey of self-play in reinforcement learning.

Haddad, G., Julia, R., e Faria, M. (2021). Técnicas de aprendizado por reforço aplicadas em jogos eletrônicos na plataforma geral do unity. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 571–582, Porto Alegre, RS, Brasil. SBC.

Huh, D. e Mohapatra, P. (2024). Multi-agent reinforcement learning: A comprehensive survey. arXiv:2312.10256.

Huynh, N.-M., Cao, H.-G., e Wu, I.-C. (2025). Multi-agent training for pommerman: Curriculum learning and population-based self-play approach. arXiv:2407.00662.

Li, Z., Ji, Q., Ling, X., e Liu, Q. (2025). A comprehensive review of multiagent reinforcement learning in video games. *IEEE Transactions on Games*, 17(4):873–892.

Milani, S., Juliani, A., Momennejad, I., Georgescu, R., Rzepecki, J., Shaw, A., Costello, G., Fang, F., Devlin, S., e Hofmann, K. (2023). Navigates like me: Understanding how people evaluate human-like ai in video games. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Proc. CHI 2023)*, pages 1–18, Hamburg, Germany. ACM.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., e Klimov, O. (2017). Proximal policy optimization algorithms.

Servat, A. e Mohamadi, H. S. (2023). Immersive game worlds: Using deep reinforcement learning for lifelike non-player characters. In *Proceedings of the 2023 International Serious Games Symposium (Proc. ISGS 2023)*, pages 1–5, Tehran, Iran. IEEE.

Sutton, R. S. e Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 2 edition.

Yannakakis, G. N. e Togelius, J. (2025). *Artificial Intelligence and Games*. Springer Nature, Cham, Switzerland.