Analyzing the Algorithms and Evaluation of Homo Sapiens Artificial Intelligence Agents in Application - A Case Study of Gaming

Yanting Yu

{yuyanting1023@outlook.com}

School of Information and Intelligent Engineering, Sanya University, Sanya, Hannan Province, 572000, China

Abstract. Traditional real-world scenarios are often costly, uncontrollable, and difficult to reproduce, while game environments have gradually become an important testbed for the research and evaluation of AI algorithms. This paper systematically reviews the application of games as complex testbeds in the research of AI agent algorithms, focusing on analyzing the testing value of different types of game environments, the current development status of mainstream AI agent algorithms, and common evaluation metrics and test frameworks.This paper focuses on studying and summarizing the application characteristics of game testbeds in AI agent research, clarifying the differentiated value of different types of game environments for algorithm testing, and clarifying the development trajectory and applicable scenarios of mainstream AI agent algorithms. At the same time, it summarizes the core challenges currently faced in this field and points out that the combination of large language models and multi-agent systems is an important future development direction.

Keywords: Game Testbeds; AI Agents; Reinforcement Learning; General Intelligence Evaluation; Multi-Agent Systems

1. Introduction

With the rapid development of artificial intelligence (AI) technology, the decision-making, learning, and collaboration capabilities of AI agents in complex environments have become a research focus. Clearly-defined rules, complicated state spaces, and the presence of repeated experiments have slowly grown to be a significant testbed to test and evaluate AI agent algorithms. The purpose of this paper is to logically review the use of games as a detailed testbed in both the study and testing of AI agent algorithms, summarize the existing research findings, and discuss the future progress trends.

Games have also gained remarkable interest as a valuable test bed and experimental system of AI research in recent years. Other scholars have extensively debated the architecture of game-based agents, assessment framework and learning algorithms in various capacities, which has formed a significant base upon which the formation of general AI and

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365242

reinforcement learning technologies build upon. The article "Games for Artificial Intelligence Research: A Review and Perspectives" goes through the systematic review of how games have been used throughout AI research, indicating that through games it is possible to test the effectiveness of algorithms in perception, decision-making and learning thoroughly in relation to controlling complexity and provides a summary of the key research directions and future prospects of game AI [1]. Generally speaking, Perez-Liebana et al. suggested General Video Game AI (GVGAI) multi-track evaluation model, and this is factored when agents, game design, and content generation algorithms are evaluated in a consistent manner. This architecture places more emphasis on the flexibility of the algorithms used in unknown game conditions, as opposed to having optimal performance on a one task basis, an essential experimental standard of general AI investigation [2]. To resolve the problem of standardized testing to reinforcement learning algorithms, Bellemare et al. developed the Arcade Learning Environment (ALE), which offered a generic evaluation environment to general agents by offering a platform of sequential playing arcades games. It has been found that this environment has successfully facilitated the creation of deep reinforcement learning algorithms (as in DQN) and formed the foundation of many of the subsequent studies [3]. In order to additional investigate the generalization and transfer potential of agents, Juliani and colleagues suggested the Obstacle Tower challenge that could combine several challenges, including visual perception, continuous control and long-term planning in a three-dimensional scene to investigate the performance of agents in the unseen levels, with the emphasis on the value of cross-task generalization in the studies of game AI [4]. Also, in "A Survey of Deep Reinforcement Learning in Video Games" developed by Arulkumaran, the systematic review of mainstream algorithms and application scenarios, the authors analyzed the pros and cons of deep reinforcement learning at the level of strategy learning, exploration mechanisms, and sample efficiency and noted that the future research should achieve a further breakthrough in the stability and generalization capabilities [5]. The studies have been performed in various ways and include such points as theoretical review, construction of the evaluation platform, testing of the generalization ability and the algorithm development, which gives enough evidence to develop more general and fundamentally cognizing game agents.

The paper seeks to examine the evaluation and learning process of the agents in terms of general game AI research, with regard to the research advancement and the application merit of AI algorithms through game setups in terms of generality and generalization powers[6]. To begin with, this paper thoroughly examines the state of the art in developing game artificial intelligence and associated evaluation systems, as well as evaluates the role and nature of game environment in research of artificial intelligence. Secondly, it provides the summary and comparison of mainstream algorithms of deep reinforcement learning and their performance in various kinds of game tasks. Lastly, it talks about the core issues the current game artificial intelligence research encounters as per the current research findings with a discussion of future research perspectives.

2. Research Analysis of AI Agents in Games as Complex Testbeds

2.1 Value Analysis of Game Environments as Testbeds for AI Agents

Game environments, with their coexistence of high structure and complexity, are widely used in the research and evaluation of artificial intelligence agent (AI Agent) algorithms. Compared with real-world environments, game testbeds have the following significant advantages:

Firstly, they have strong controllability and repeatability. The rules, state transitions, and reward mechanisms of game environments are definite, allowing researchers to repeat experiments under the same conditions multiple times, thus ensuring the comparability and scientific nature of the experimental results.

Secondly, the state space and task complexity can be adjusted. From low-dimensional pixel input environments like Atari to open-world and real-time strategy games, game environments can gradually increase the difficulty of perception, decision-making, and planning, making them suitable for algorithm verification at different stages[7].

Thirdly, they are low-cost and highly safe. The testing of robots or autonomous driving in the real world is costly and poses safety risks, while game environments can verify algorithm performance without risks.

Therefore, games have gradually evolved from "entertainment systems" to standardized experimental platforms in artificial intelligence research, serving as an important bridge connecting theoretical research and real-world applications.

2.2 Classification and Characteristics of the Various Game Testbeds.

Based on the description of the task structure and information features, the game testbeds that are regularly used can be roughly grouped into a few categories. Playing classic board games (Go, chess, Texas Hold'em, etc.) allows understanding all the rules and fairly manageable state space, which makes them the right object of search algorithm and game theory research. General video game environments (as e.g. ALE and GVGAI) have numerous different-rule games, with cross-task learning and generalization capacity, and such environments are also significant in the measurement of general intelligence. Visual perception, continuous control, and long-term planning aids are more required in three-dimensional visual and control environments (Obstacle Tower, Minecraft, and Unity ML-Agents). Multi-agent collision and competition games offer, meanwhile, natural experimental arenas to study the mechanism of communication and collaboration strategies and games evolution [8].

The various categories of game testbeds have complexity and evaluation priorities towards which a multi-layer experimental system of AI agent-research can be constructed.

2.3 Different Types of Games Can Be Mainly Categorized into Four Types

For example, classic board games include Go, chess, and Texas Hold'em. These games feature well-defined rules and enumerable states, making them suitable for studying search algorithms, game theory, and strategic optimization problems. The success of AlphaGo and AI in Texas Hold'em has demonstrated the effectiveness of combining reinforcement learning with search. For instance, the Universal Video Game Environment [9].

The category of platforms is the one represented by the Atari Learning Environment (ALE) and General Video Game AI (GVGAI). It consists of a huge number of all kinds of games, all with different rules, which are used to test the learning and generalization abilities of AI agents in general, and thus it is a contemporary field of research. As an example, 3D vision and control task settings include such platforms as Obstacle tower, Minecraft and Unity ML-Agents. These conditions also place more emphasis on the capabilities of visual perception, continuous control, and long-term planning, and they have an increased requirement on the strength and generality of algorithms. Multi-agent games Multi-agent games, also called cooperative/competitive games, are designed to investigate multi-agent cooperation, communication, and the dynamics of the development of game-theoretical strategies. MOBA games, and multiplayer adversarial environments are examples of the latter.

2.4 AI Agent Algorithms: Game Testbeds.

In game testbeds, AI agent algorithms have rapidly evolved from rule-based and search-based methods to deep reinforcement learning and its extended forms. Early methods relied on manual rules and search strategies, which were highly interpretable but had limited generalization capabilities. Deep reinforcement learning, by introducing deep neural networks, enables agents to learn strategies directly from high-dimensional perceptual inputs, achieving breakthroughs in Atari and 3D game environments. Further, multi-agent reinforcement learning focuses on strategic games in collaborative and competitive scenarios, while the recent rise of "large language models + agents" research attempts to utilize language models for task understanding, strategy planning, and decision explanation, providing new ideas for enhancing generality and interpretability[8].

3 Data Analysis and Evaluation Methods Review

3.1 Common Evaluation Metrics for Game AI Agents

In game testbeds, researchers typically evaluate AI from several dimensions, as shown in Table 1:

Table 1 Dimensional Analysis

Evaluation dimensions	Common indicators
performance	Average score, win rate, task completion rate
learning efficiency	Convergence speed and sample complexity
generalization ability	No performance under the new rules of the level has been observed
stability	Variance of multiple experimental results
robustness	Performance changes under noise or environmental disturbances

3.2 Comparative analysis of evaluation data from different game testing platforms

Regarding the Atari platform, Bellemare et al. proposed the Arcade Learning Environment (ALE), which focuses on classic arcade games with fixed rules. The platform mainly measures the learning effectiveness and strategy optimization ability of agents in a single task environment through game scores, and is suitable for evaluating the basic learning performance and stability of reinforcement learning algorithms [3].

In terms of GVGAI platform, the General Video Game AI framework proposed by Perez Liebana et al. emphasizes real-time adaptability in various unknown games, and its core evaluation objective is not a single game score, but the generalization ability and robustness of the agent in cross game scenarios. It is widely regarded as an important experimental platform for testing "general intelligence" [2].

In response to higher-level generalization and transfer problems, Juliani et al. proposed the Obstacle Tower environment, which generates level structures through programmatic random generation, introduces visual changes, continuous control, and long-term planning tasks, and focuses on examining the visual understanding and long-term decision-making abilities of agents in unseen environments, thus compensating for the shortcomings of traditional 2D gaming platforms in generalization evaluation [4].

Based on the conclusions of multiple studies, Arulkumaran et al. pointed out in their review that algorithms that achieve excellent performance in a single gaming environment often have difficulty maintaining performance advantages in complex or completely new environments. This phenomenon reflects that current gaming AI agents still have significant shortcomings in generalization and transfer abilities, and has become one of the key issues restricting the development of general artificial intelligence [5].

3.3 Main difficulties faced by current research

Although game test beds have achieved significant results in AI research, there are still several key challenges. Firstly, the high performance achieved by algorithms in a single environment is often difficult to transfer to new tasks or rules, and insufficient generalization ability remains the core bottleneck. Secondly, deep reinforcement learning methods are highly dependent on data and computing resources, have high training costs, and lack interpretability in the learning process. In addition, the inconsistent testing environments and evaluation indicators used in different studies also affect the comparability of research results to a certain extent.

3.4 Future research directions and prospects

Future research can be advanced from multiple directions. One is to build a more diverse and standardized game testing bed to systematically evaluate the general intelligence level of AI agents. Secondly, explore multi task learning, meta learning, and self supervised learning methods to enhance the rapid adaptation ability of agents in unknown environments. Thirdly, combining big language models with multi-agent systems is expected to achieve breakthroughs in task understanding, strategy generation, and collaborative decision-making. Through in-depth research in these areas, game test beds are expected to play a more important role in building a more universal and reliable artificial intelligence agent evaluation system.

4 Conclusion

This article takes the game environment as a complex testing bed as the research object, and adopts the methods of literature review and comparative analysis to systematically sort out and summarize the algorithm application, evaluation system, and research progress of artificial intelligence agents in different types of game environments. Through comprehensive analysis of classic game genres, general video game environments, and 3D vision and control task platforms, the unique value of game test beds in evaluating AI agent learning ability, decision-making ability, and generalization ability is explored.

The research results indicate that there are significant differences in the evaluation focus among different game testing platforms. Platforms represented by Atari Learning Environment focus more on improving the learning efficiency and performance of algorithms under a single task, and are suitable for verifying the fundamental capabilities of deep reinforcement learning algorithms; General video game platforms such as GVGAI emphasize the adaptability of agents in multitasking and unknown rule environments, which is an important tool for measuring the level of general intelligence; 3D environments such as the Obstacle Tower require higher visual understanding, continuous control, and long-term planning abilities from agents through randomly generated levels. Existing research has generally found that algorithms that perform well in a single gaming environment often struggle to maintain stable performance in complex or completely new environments, exposing the shortcomings of current AI agents in terms of generalization ability, robustness, and interpretability[10].

In addition, this article also summarizes the main challenges facing current research, including the high dependence of deep reinforcement learning on data and computing resources, the lack of unified standards for evaluation metrics and testing environments, and the difficulty in explaining algorithm decision-making processes. These factors have to some extent constrained the transfer of game AI research results to real-world applications.

Looking ahead, game test beds will continue to play an important role in artificial intelligence research. On the one hand, by building a more diverse and standardized testing environment, the general intelligence level of AI agents can be systematically evaluated; On the other hand, combining new methods such as multi task learning, meta learning, and large language models is expected to enhance agents' ability to quickly adapt and interpret decisions in unknown environments. Related research is not only of great significance for promoting the development of general artificial intelligence, but also provides valuable references for the secure application of artificial intelligence in complex real-world scenarios.

References

[1] C. Hu, Y. Zhao, Z. Wang, H. Du and J. Liu,“Games for Artificial Intelligence Research: A Review and Perspectives,”IEEE Transactions on Artificial Intelligence,vol. 5, no. 12, pp. 5949–5968, Dec. 2024.

[2] D. Perez-Liebana, S. Samothrakis, J. Togelius, T. Schaul and S. Lucas,“General Video Game AI: A Multi-Track Framework for Evaluating Agents, Games and Content Generation Algorithms,”IEEE Transactions on Computational Intelligence and AI in Games,vol. 8, no. 3, pp. 195–214, Sept. 2016.

[3] M. G. Bellemare, Y. Naddaf, J. Veness and M. Bowling,“The Arcade Learning Environment: An Evaluation Platform for General Agents,”Journal of Artificial Intelligence Research,vol. 47, pp. 253–279, 2013.
[4] A. Juliani, A. Khalifa, V. Berges, J. Harper, E. Teng, et al.,“Obstacle Tower: A Generalization Challenge in Vision, Control, and Planning,”in Proceedings of the AAAI Conference on Artificial Intelligence,vol. 33, no. 1, pp. 2684–2691, 2019.
[5] K. Arulkumaran, M. P. Deisenroth, M. Brundage and A. A. Bharath,“A Survey of Deep Reinforcement Learning in Video Games,”IEEE Transactions on Computational Intelligence and AI in Games,vol. 9, no. 4, pp. 414–427, Dec. 2017.
[6] J. Pfau, J. Togelius and A. Khalifa,“Automated Video Game Testing Using Synthetic and Human-Like Agents,”in Proceedings of the IEEE Conference on Games (CoG),pp. 1–8, 2021.
[7] S. Risi and J. Togelius,“Augmenting Automated Game Testing with Deep Reinforcement Learning,”in Proceedings of the IEEE Conference on Games (CoG),pp. 1–8, 2020.
[8] Zhang, H, Research on the Explainability of AI Strategies for Tetris Games Based on Large Language Models Zhang Haonian Research on the Explainability of AI Strategies for Tetris Games Based on Large Language Models [D]. Nanjing University of Information Science and Technology, 2026. DOI: 10.27251/d.cnki.gnjdc.2024.000673.
[9] Yuan W, Liao, Z, Gao, W, Wei, T., Luo J, Zhang W, Chen J, “Review of Computer Poker Intelligent Game Playing (Related to Texas Hold'em Poker)“. Journal of Network and Information Security, 2021, 7(5): 57–76.
[10] J. McCarthy and M. Minsky,“The Role of Artificial Intelligence in Gaming,”AI Magazine,vol. 39, no. 2, pp. 15–28, 2018.