Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.34293

Research and Analysis on the Interaction Mechanism, System Architecture, and Alignment Issues of Generative Intelligent NPCs

Hongbo Wang

College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, Beijing, China
wonghongbo1@gmail.com

Abstract. With the development of large language models (LLMs) and generative artificial intelligence, non-player characters (NPCs) are shifting from traditional rule-driven systems to generation-driven systems. Traditional NPCs mainly rely on finite state machines and behavior trees to implement predefined behavioral logic. Although these methods provide strong stability and controllability, they cannot fully meet the demand for high freedom, strong immersion, and natural interaction in open-world games. Generative intelligent NPCs introduce memory, reasoning, planning, and natural language understanding capabilities, allowing game characters to evolve into agents with persistent interaction abilities. This paper examines the evolutionary history, interaction loops, system architecture, functional paradigms, industrial implementation, and alignment challenges of generative intelligent NPCs. Drawing on the theory of the digital labyrinth and wax wings, the Generative Agents framework, LLM-based agent architecture, and industrial cases such as NVIDIA ACE and NetEase Fuxi, this paper analyzes the application paths of generative NPCs in emotional companionship, strategic opposition, and real-time interaction. The study argues that generative NPCs are transforming game interaction from script-driven content delivery to experience-driven world building. However, issues such as behavioral loss of control, value drift, computational cost, and ethical governance still need further attention. In the future, balancing intelligence, real-time response, and controllability will become the core direction for the sustainable development of generative NPCs.

Keywords: Generative artificial intelligence, non-player character, large language model, agent, AI alignment

1. Introduction

Non-player characters (NPCs) are among the most basic and important interactive entities in digital game worlds. They have long undertaken multiple functions such as mission guidance, plot advancement, combat confrontation, and emotional companionship. The intelligence of the NPCs directly affects the players' sense ofRealism, Immersion and Participation in Virtual Worlds. Early on Reactive characters based on finite state machines (FSMs), to further complex decision-making

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

DOI: 10.54254/2755-2721/2026.34293

characters are implemented using behaviour trees (BTs), and then to generative agents based on large language models (LLMs), Development of NPCs is also related to the changes in artificial intelligence in games. Intelligence, and the transformation of digital interaction systems. Script Control and Autonomous Generation.

The first generation of NPCs are rule-based. Player Behaviour. FSMs have disjointed blocks of behaviour. States include patrol, chase and attack, and switch between these states. based on the above conditions. Behavior Trees add rules for behaviour. Organisation by a top-down hierarchy. However, these systems still rely on the developers' previous listing of behavioural paths. As a result, They cannot address the problems of complex, dynamic and unpredictable user behaviour. in Open-world Environments. The aforementioned studies have indicated that FSMs and BTs laid the engineering foundation for game AI, but they also have obvious Deficiencies in long-term memory, contextual awareness, and dynamism. Adaptation [1, 2].

With the development of generative AI, a large number of applications have gradually emerged. LLMs for natural language understanding, reasoning and planning, etc. Memory Management: NPC systems have moved away from scripts and are now more dynamic. Generation-driven systems. Park and others proposed the Generative Agents. framework that can help virtual characters manage their daily lives, spread information and form social networks through history. Experience via memory streams, reflection and planning modules [3]. Now NPCs are no longer mere vehicles for running fixed scripts.

Li Dianfeng Describes the Development of Generative Intelligent NPCs Using the metaphor of the digital labyrinth and wax wings. The first one Hands are large models that can give NPCs more autonomy and generation capabilities. capability and, therefore, to break out of the constraints of traditional Scripts featuring characters with wax wings. On the other hand, this Freedom also leads to a system that is complicated and difficult to control. Digital labyrinth, and the pathways of behaviour are no longer entirely predictable. And the risk of misalignment, safety and ethics will also be relatively high [4]. Wang and others have put forward the following four modules for LLM agents. Profile, Memory, Planning and Action are the foundations of this. To analyze the structure of the generative NPC system [5].

Based on the above developments, this paper examines the interaction Mechanisms, System Architecture and Alignment Issues of Generative Models Intelligent NPCs. It constructs the following system. Evolutionary History, Interaction Loops, Functional Paradigms, Industrial Implementation and Problems and Prospects. A combination of the above. literature review and case analysis to summarize the Development Path and Key Technical Logic of Generative NPCs. Future Applications and Governance Paths for These in the Game Industry.

2. Development history of NPCs

The development of NPCs in digital games is a record of iteration. Computer algorithms and the transformation of artificial intelligence. Intelligence Logic from Simple Reaction to Human-like Cognition. In NPC development has gone through the three general stages of finite-state machines. Machines, Behavior Trees and Generative Agents.

2.1. Early logic and the rise of finite state machines

Early NPCs in video games were relatively simple interactive objects. To For instance, the opponent's logic in Pong mainly used physical collision. Position judgement and lack of willingness to move. Afterwards, a finite-state machine served as the first kind of NPC design. A state machine

DOI: 10.54254/2755-2721/2026.34293

has several states and transitions among them for the NPC. via pre-defined conditions to have the characters do simple Behaviour of patrol, pursuit and attack. Their Advantages have a good structure, are easy to implement and stable. However, as the game scenes become more complex, the number of states also increases. increased, leading to a reduction in service life and deformation.

2.2. Hierarchical decision-making and the maturity of behavior trees

After the 2000s, behaviour trees have gradually become widely used in NPC control. Schemes in the game industry. Behavior trees are less complex than FSMs. behavior logic through selectors, sequence nodes, condition nodes, etc. Action nodes: Improving Modularity and Scalability. Many combat, stealth, and open-world games employ behaviour trees for enemy patrol and alert systems. Search and Attack Logic Behavior trees improve NPC strategy and environmental awareness, but their decision space is still mainly predefined by developers. Therefore, they cannot fully respond to nonlinear behavior generated by players in open-world environments.

2.3. The rise of generative agents and quasi-subjectivity

In the 2020s, the introduction of LLMs made it possible for NPCs to move from rule executors to generative agents. Generative Agents uses memory streams, reflection, and planning mechanisms to allow virtual characters to generate schedules, maintain relationships, and produce continuous behavior based on their experiences [3]. Research on AI-generated content also shows that generative technologies reduce the threshold of game content production, enabling character dialogue, mission text, and scene feedback to be generated in real time according to context [6]. NPC control logic is developing step-by-step in evolution. Shift from human-in-the-loop to human-out-of-the-loop. Interaction Feedback no longer needs to be in the form of an all-encompassing script. More choices of intelligent decision-making [4].

3. Interaction loop and system architecture

The center of generative intelligent NPCs is no longer script execution. in the traditional sense, but a continuous interaction loop composed of Perception, cognition, memory, reasoning and action. Li Dianfeng The concept of a digital labyrinth and wax wings can explain this change. LLMs provide more autonomous generation for NPCs on the one hand. capabilities to break out of the fixed path of FSMs and Behaviour Trees. At the same time, the stronger the generation ability Capability refers to the difficulty in exhausting all possible behaviour paths. and the harder it is to explain and control [4].

Generative NPCs are introduced to study interaction mechanisms. be divided into three kinds: human-in-the-loop, human-on-the-loop, and unsupervised learning. loop and human-in-the-loop. Human-in-the-loop promotes real-time Human supervision of NPC behaviour and is suitable for key plot nodes and High-risk Interactions. Human-on-the-loop enables NPCs to be semi-autonomous. decisions within the rules, and people are mainly responsible for Rule Design and Anomaly Correction. Human-in-the-loop further Increase NPC autonomy and enable NPCs to use environment perception. and reasoning mechanisms for behaviour planning. These three modes Show the evolution of control priority to autonomy priority. and correspond to the structural tension between intelligence and Controllability [3, 4].

Wang and others have put forward the system architecture of LLMs. The four modules of the agents are: Profile, Memory, Planning and Action [5]. Profile is the ID and traits of a non-player

DOI: 10.54254/2755-2721/2026.34293

character. Memory stores past experiences, world knowledge and long-term memories. Relationships. Planning is responsible for goal decomposition and task. Ordering and Strategy Selection. Action translates model predictions into Dialogue, Movement, Combat or Social Behaviour. It will be seen here. Paper abstracts present generative NPCs as follows:

$$\text{NPC} = (\text{Perception}, \text{Memory}, \text{Decision}, \text{Action}) \tag{1}$$

$$\text{Action} = \text{f}(\text{Input}, \text{Memory}, \text{Persona}) \tag{2}$$

Input is the player's current input and the environment. State and Memory refer to past interactions and extended experiences, respectively. Persona represents a person's identity, values and behaviour. Style. The model shows that NPC behaviour is not only a result of the present. instructions, as well as in the long-term memory and character constraints.

Park and others have shown that this is practical in fact. Long-term NPC behaviour simulation based on a memory stream-reflection mechanism. Planning Mechanism [3]. However, in industrial deployment, a full Dependence on LLMs may have a longer response time and unpredictable output. High operating expenses. Thus, the present solution is relatively feasible. Hybrid architecture of LLMs and Behavior Trees/FSMs. The old rule system is suitable for a short period of high-frequency operation. attack, and pathfinding; LLMs handle semantic understanding, but not tasks. Planning and Natural Language Generation. The Structure maintains Stability of Traditional Game Systems and Creativity in New Works. Generative models are also one of the main ways to do this now. Implementation of Generative NPCs.

4. Functional paradigm: rope-stick design for generative NPCs

Li Dianfeng has proposed a functional design for generative Intelligent NPCs fall into two categories: ropes and sticks [4]. A rope Paradigm: Focus on emotion, companionship and story. guidance to help players experience life in the game. The rod Paradigms promote competition, challenge and restrain behaviour. To generate pressure and increase the strategic depth. These two Paradigms are not mutually exclusive but rather form a new system. Balance in Different Game Types and Interaction Scenarios.

Rope-type NPCs are generally classified as emotional, companion or narrative. characters. NetEase Fuxi's AI teammate enhances Companion Experience in Multiplayer Competitive and Open-World Environments Real-time Dialogue, Intention Recognition and Contextual Memory [7]. The Intelligent NPCs in Justice Mobile have followed a similar trend: characters can understand the players' natural language input and build a system. Relationship memories from long-term interaction. Such Platforms include: Inworld AI and Convai offer overseas developers NPC solutions. with a long-term persona. The above solutions are commonly used to Tavern owners, mentors, companions, and mission characters in the role-playing game. Playing Games [8]. The value of such NPCs is not to increase the game Difficulty, but in enhancing emotional feedback and character credibility. and virtual realities.

Stick-type NPCs are generally competitive in nature and high-pressure games. play and are frequently seen in boss fights, PvE enemies, and tactical Confrontation Scenarios. Traditional Behavior Trees can Design Phases based attacks, patrol alerts and tactical responses; reinforcements Learning tools are available to train a more adaptive adversarial network. Characters [9]. Bosses in Souls-like games are often depicted as being. Pressure expressions in terms of phase transitions, attack rhythms and behaviour. Patterns. Enemies in tactical shooting games help develop players'

DOI: 10.54254/2755-2721/2026.34293

sense of Flanking, Cover Use and Resource Competition as Defenses. The second LLMs will be applied to the top-level planning work in the future. Enable the adversarial NPCs to have better environmental awareness. and strategic modification.

Many NPCs in modern games have combined ropes with other forms. Stick attributes. For example, the main support characters in a work. Narrative companionship and emotional communication, as well as supporting. Combat support for missions. Generative NPCs are therefore somewhat problematic. moving from single-function design to composite functionality: They must make players want to approach them and also make them It is requested. Both kinds of NPCs share the same system. Architectures of perception, memory, planning and action, but their Optimization Objectives Differ. Rope-type NPCs are more emotionally evocative. and extended connections; stick-type NPCs are focused on strategy. Performance and behavioural Stability.

5. Technical route and industrial application

Generative NPCs in industry applications are not just new functions. Matter of Naming a Large Language Model. It is a collaboration between. Several modules, including speech recognition and language reasoning. Character memory, action driving, safety filtering, and automation. Testing. NVIDIA ACE is a typical industrial application. Using Automatic Speech Recognition (ASR) and small language models (SLMs). and digital human driving modules to realise real-time interaction. Chain from hearing to understanding, thinking and expressing [10]. Generative NPCs also require the full set of system engineering. Support, but not a single chat model.

Figueiredo and others suggest that character consistency be achieved through Structured Prompts, Character Constraints and Retrieval-Augmented Generation. Generation mechanism for improving the stability of NPC dialogue [11]. It The other is the phenomenon of changes in personality and stories. Due to free generation. Character Identity and Task Division are embedded. Emotional Style and World Knowledge have been added to the symbolic prompt structure. allowing NPCs to speak naturally and not easily diverge from this Their Settings.

NetEase Fuxi in China is focused on later-stage optimization. Training and Scenario-based Reasoning Abilities. Its AI assistant System uses intention recognition, chain-of-thought reasoning and character profiling mechanisms for NPCs to understand better Players' explicit instructions, as well as the goal they wish to achieve implicitly. Tactical Requirements [7]. NPCs are usually unable to do the same. Based on the Battle Situation, decide whether the player needs supplies. covering, or a coordinated attack, rather than being in the same class as Superficial Question Answering.

With an increase in the complexity of generative NPCs, so too will be the necessary quality assurance. Also have been used in industry to some extent. Neelapu's Work Automated testing for AI-generated game content and using the technology. of generative test scripts, anomaly detection and end-to-end verification mechanisms for logical deviation and narrative discontinuity. Improper Output [12]. Generative NPCs are now broadly in development. Single-point Dialogue Generation for Full-Chain Intelligent Agents. System The front end is based on speech and multimodal perception, etc. The middle layer has prompts, memory and planning, and the back-end tests, filters and alignments to ensure the following Stability.

6. Challenges and future prospects

As generative NPCs are increasingly applied in open-world, social companionship, and tactical confrontation scenarios, alignment issues have become a key research concern. Ji et al. point out that AI systems need to balance robustness, interpretability, controllability, and ethicality [13]. This

DOI: 10.54254/2755-2721/2026.34293

paper summarizes these dimensions as the RICE principle. Robustness emphasizes stable performance under abnormal input and complex environments. Interpretability requires developers to understand the logic of NPC behavior generation. Controllability focuses on maintaining boundary constraints even under high autonomy. Ethicality requires NPC behavior to conform to basic social norms and value consensus. Gabriel's research on value alignment also suggests that AI systems should consider human values and social norms beyond technical goals [14].

In practical applications, generative NPCs may face multiple risks. First, long-term memory accumulation, context drift, or prompt injection may cause behavioral deviation, leading to character personality distortion and even damage to narrative structure. Second, companion NPCs may create emotional manipulation risks due to excessive anthropomorphism, especially in romance, companionship, and games for minors. Third, in multi-agent environments, autonomous collaboration among NPCs may produce unpredictable emergent behavior, increasing the difficulty of debugging and supervision. Research on LLM agent alignment further points out that traditional output filtering is insufficient to cover the full decision-making process of agents. Future systems will need to combine reinforcement learning from human feedback, Constitutional AI, scalable oversight, and multi-agent governance mechanisms to achieve process-level constraints [15].

In addition to safety and ethical concerns, computational resources and real-time response capability are also core bottlenecks for the large-scale implementation of generative NPCs. Although LLMs have strong reasoning and generation capabilities, their inference cost is high and their response latency is often long. They are therefore difficult to use directly in combat systems, multiplayer competitive scenes, and high-frequency interactions that require millisecond-level feedback. A more realistic solution is the hybrid architecture of LLMs plus predefined behavior patterns. Behavior trees, FSMs, or reinforcement learning modules are responsible for high-frequency actions such as movement, attack, defense, and pathfinding, while LLMs are responsible for high-level semantic understanding, emotional interaction, and complex planning. With the development of edge inference, small model distillation, local deployment optimization, and dedicated AI chips, the real-time response capability of generative NPCs is expected to improve. Nevertheless, hybrid architecture will remain the mainstream path in the foreseeable future.

Future research can proceed in three directions. The first is long-term memory and persona consistency, which can improve the stability of NPCs during long-cycle interaction. The second is alignment mechanisms for agent systems, including behavior verification, interpretable reasoning, and safety constraints. The third is multi-agent collaboration and social simulation, enabling NPCs to form more realistic group behavior in complex worlds. At the same time, the improvement of industry standards, copyright rules, and ethical review mechanisms will also become an important foundation for the healthy development of generative NPCs.

7. Conclusion

This paper studies the interaction mechanisms, system architecture, and alignment issues of generative intelligent NPCs. It systematically reviews the development path of NPCs from finite state machines and behavior trees to LLM-driven generative agents, and analyzes the key logic of interaction loops, functional paradigms, and industrial implementation through cases such as Generative Agents, NVIDIA ACE, and NetEase Fuxi. The study shows that generative NPCs are transforming game interaction from script-driven systems to experience-driven systems. NPCs are no longer merely tool characters for mission delivery and combat confrontation, but are gradually evolving into generative agents with memory, planning, and persistent interaction capabilities.

DOI: 10.54254/2755-2721/2026.34293

At the same time, this paper argues, based on the theory of the digital labyrinth and wax wings, that the development of generative NPCs is essentially a search for balance between intelligence and controllability. In the future, with the continuous improvement of long-term memory, multi-agent collaboration, edge inference, and alignment mechanisms, NPCs will further participate in the dynamic construction of game worlds. However, issues such as computational cost, behavioral loss of control, and ethical governance still require continuous attention. Only through both technological innovation and institutional regulation can generative NPCs achieve healthy and sustainable development.

References

[1] Biggar, O., Zamani, M., & Shames, I. (2021). An expressiveness hierarchy of behavior trees and related architectures. IEEE Robotics and Automation Letters, 6(3), 5397-5404. https: //doi.org/10.1109/LRA.2021.3074337

[2] Assaf, M. (2023, December 15). From Pong to narrative: The evolution of AI in gaming. ART 108: Introduction to Games Studies, San Jose State University ScholarWorks. https: //scholarworks.sjsu.edu/art108/37/

[3] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), Article 2. ACM. https: //doi.org/10.1145/3586183.3606763

[4] Li, D. (2024). Digital labyrinth and wax wings: Prospects and challenges of generative intelligent NPCs in video games. Journal of Beijing Film Academy, (12). (Translated from Chinese)

[5] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345. https: //doi.org/10.1007/s11704-024-40231-1

[6] Wu, Y. (2025). Innovative applications of AIGC technology in game narratives. Journal of Modern Social Sciences, 2(4), 287-292. https: //doi.org/10.71113/JMSS.v2i4.350

[7] NetEase Fuxi. (2025, October 11). When game NPCs have a soul: NetEase Fuxi decodes new practices in intelligent game interaction scenarios. Volcengine ADG Community. (Translated from Chinese). https: //adg.csdn.net/6970a470437a6b40336b0483.html

[8] Inworld AI; Convai. (2025). AI NPC platform documentation and cases. Inworld AI and Convai. https: //inworld.ai/; https: //www.convai.com/

[9] Unity Technologies. (2023). Unity ML-Agents Toolkit. GitHub. https: //github.com/Unity-Technologies/ml-agents

[10] NVIDIA. (2025, February 20). Bring NVIDIA ACE AI characters to games with the new In-Game Inferencing SDK. NVIDIA Technical Blog. https: //developer.nvidia.com/blog/bring-nvidia-ace-ai-characters-to-games-with-the-new-in-game-inference-sdk/

[11] Figueiredo, V., & Elumeze, D. (2025). Symbolically scaffolded play: Designing role-sensitive prompts for generative NPC dialogue. arXiv preprint arXiv: 2510.25820. https: //arxiv.org/abs/2510.25820

[12] Neelapu, M. (2025). Automated QA testing for AI-generated game content: Using LLMs to validate NPC behavior and narrative integrity. International Journal of Emerging Trends in Computer Science and Information Technology (IJETCSIT), 198-208.

[13] Ji, Z., Liu, Z., Lee, N., Yu, T., Wilie, B., Zeng, M., et al. (2024). AI alignment: A comprehensive survey. arXiv preprint arXiv: 2310.19852. https: //arxiv.org/abs/2310.19852

[14] Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437. https: //doi.org/10.1007/s11023-020-09539-2

[15] Zhou, D., Zhang, J., Feng, T., & Sun, Y. (2025). A survey on alignment for large language model agents. UIUC Spring 2025 CS598 LLM Agent Workshop. https: //openreview.net/forum?id=gkxt5kZS84