Behavior is all you need

Vishnu Hari[*] Connor Brennan[*]

Abstract

Current game bots lack the emotional depth, spontaneity, and intent that make human players compelling, resulting in predictable interactions and broken immersion. We propose a framework for building believable AI agents that behave convincingly enough to evoke emotional engagement without requiring true sentience. Drawing from narrative theory and human psychology, our approach focuses on four key components: personality, needs, emotion, and memory. Together, these modules drive behavior that appears motivated, context-sensitive, and consistent over time. Much like how audiences connect with fictional characters, players can suspend disbelief when bots behave in ways that suggest underlying intention. This play-acting approach offers a scalable path to lifelike game agents — enabling emergent gameplay, dynamic world population, automated testing, and emotionally resonant interactions — not by replicating consciousness, but by engineering behavior that is rich enough, reactive enough, and real enough to be believed.

Contents

[*] Lead authors. Correspondence should be directed to v@ego.live.

1 Introduction

Humans in multiplayer games offer a fundamentally richer and more engaging experience than bots, due to their unpredictability, emotional reactivity, and complex intentions [Bates, 1994, Nass and Moon, 2000]. Anecdotal examples illustrate this clearly: whether it is summoning a mech in Helldivers 2 and accidentally flattening your friends, or hearing a colleague shouting in frustration during a competitive FPS session, these events are emotionally resonant because the participants are human. They respond in nuanced, emotionally driven, and sometimes irrational ways that bots cannot yet replicate. Bots tend to behave in predictable, mechanical patterns that lack spontaneity or emotional variability, causing interactions to feel flat and forgettable by comparison.

One core reason for this disparity is the presence of perceived intentionality and emotional stakes in human play. When we compete against other players, we implicitly understand that they have goals, preferences, and internal motivations - features that we attribute to minds similar to our own [Turing, 1950, Heider and Simmel, 1944]. Thus, victory in a competitive game against a human opponent feels deeply satisfying because we have asserted our will over another conscious entity. This creates a sense of consequence and emotional payoff that bots fail to deliver. Their behavior lacks internal narrative coherence; they don't want anything, so defeating them is less impactful. Even humorous or chaotic moments in multiplayer games rely on a shared understanding of why someone acted the way they did. This understanding hinges on our ability to empathize with and interpret others' actions through the lens of human-like intention [Mateas, 1999].

To recreate this sense of meaningful interaction with nonhuman agents, we argue for a design approach based on 'play-acting' rather than true artificial sentience. In books and films, audiences routinely suspend disbelief and emotionally invest in characters they know to be fictional. This is possible because those characters exhibit believable motivations and consistent internal logic [Smith et al., 2023]. Similarly, bots in games need not be fully autonomous or sentient to feel human-like—they must simply behave in ways that allow players to intuitively project intention and emotion onto them. This means crafting bot behaviors that respond to context, reflect plausible motivations, and display transient emotional modulation [Park et al., 2023]. When bots exhibit behavior that players can interpret as emotionally and cognitively coherent, even if artificial, the illusion of humanity is preserved, and with it the emotional depth that makes games memorable.

2 Definitions

2.1 What does it mean for a bot to have human-like behavior?

Here, we define human-like behavior in the context of interactive game agents as behavior that appears intentional, emotionally grounded, and mentally coherent to human players. This does not imply true consciousness or cognitive realism. Rather, the focus is on believability: the degree to which an agent's actions can be interpreted as stemming from internal motivations such as goals, preferences, or emotions. The objective is not to model human cognition with fidelity, but to simulate behavior that feels socially and emotionally coherent [McCrae and John, 1992, Ortony et al., 1988].

This is the same principle that allows audiences to connect with fictional characters in books or films. Although we know that they are not real, we still empathize with them because their actions make sense within a recognizable human framework [Mateas, 1999]. It is the role of authors and directors to ensure that these characters feel believable. In our case, this believability will emerge automatically by ensuring that all agents exhibit behavior consistent with a defined personality and responsive to changing context. The result is not true sentience, but the illusion of agency: agents that act as if they have minds and internal lives of their own.

2.2 Current bots' mechanisms of failure

Modern bots consistently fail to meet the behavioral threshold required for immersion. Most are built around deterministic scripts or reactive state machines that result in shallow and repetitive behavior. Their dialogue is often generic, lacking any personal context or emotional variability, and their responses are typically disconnected from prior interactions [Warpefelt and Verhagen, 2017].

This problem persists even with newer LLM-driven bots. While their surface fluency can mask it temporarily, these bots are ultimately just sophisticated skins over the same ”helpful assistant” paradigm. At their core, they are fine-tuned foundation models trained to assist users in a call-and-response format — designed to reflect and validate the user's input rather than simulate an independent mind [Zhu et al., 2024]. As such, they exhibit no true intentionality: they wait to be prompted, then offer helpful, polite, and often overly agreeable replies. Even when dropped into game worlds or interactive narratives, they remain fundamentally passive, more akin to sounding boards than characters.

A key failure mode underlying both traditional and LLM-based bots is the absence of autonomy. Bots rarely initiate, pursue goals, or act in ways that suggest independent thought. Emotional flatness compounds the issue — bots often fail to respond meaningfully to praise, insult, conflict, or emotionally charged events that would elicit clear reactions from a human [Smith et al., 2023]. Moreover, most are stateless, unable to remember past exchanges or evolve their behavior over time. This lack of continuity and internal growth further breaks the illusion of believability, making the bot's artificial nature quickly apparent. Until bots are equipped with a behavioral model capable of supporting autonomous, emotionally grounded, and persistent decision-making, they will remain tools — not characters. But when this agency is added, even simple bots can begin to cross the threshold into fully human-like agents [Xie et al., 2024, del Rio-Chanona et al., 2024].

2.3 What are the minimum requirements?

To cross the threshold into believability, an agent must satisfy a minimum set of behavioral criteria:

- **Stable, Recognizable Personality:** Agents should exhibit a consistent identity that modulates decision-making and responses. This personality must persist over time, while remaining flexible enough to adapt to changing environmental or social context.

- **Internal and External Motivations (Needs):** Behavior should be driven by explicitly modeled needs such as hunger, belonging, curiosity, or personal success. These drives provide psychological coherence and help explain *why* an agent behaves the way it does [Maslow, 1943]. In multi-agent simulations social drives also play a large role and have been studied in the context of economic games [Xie et al., 2024].

- **Emotionally Modulated Behavior:** Agents should be influenced by transient emotional states that arise in response to events or stimuli. These emotional shifts introduce short-term behavioral variability and enhance perceived emotional depth.

- **Short- and Long-Term Memory:** Agents must retain both immediate situational context (short-term memory) and accumulated interaction history (long-term memory). This allows behavior to evolve meaningfully over time and prevents the impression of stateless, mechanical responses.

Together, these components enable agents to move beyond surface-level mimicry and begin approximating the expressive and adaptive nature of human behavior in games.

3 Potential Benefits of Human-like Agents

3.1 More authentic gameplay

Human-like agents have the potential to dramatically increase the authenticity of gameplay experiences. In current multiplayer games, the richness of the experience often derives from the unpredictability, emotion, and intentionality of human opponents or allies [Nass and Moon, 2000]. Bots that behave in flat, mechanical ways cannot replicate this depth. By contrast, agents with distinct personalities, motivations, and emotional reactivity can contribute to emergent moments that feel genuinely meaningful [Bates, 1994].

For example, an agent that panics and flees when outnumbered, or seeks revenge after being wronged, creates situations that feel narratively and emotionally charged [Park et al., 2023]. These moments are what players remember and talk about — often more so than scripted set pieces. Human-like agents can thus preserve the emotional weight of interactions, making victories feel earned, betrayals feel personal, and cooperation feel authentic.

3.2 Automatic population of game worlds

Populating large-scale game worlds with believable characters is a time-consuming and resource-intensive task. Designers must write dialogue, design behavior trees, and manually script events to simulate life. Human-like agents offer an alternative: agents that autonomously populate and animate the world through emergent, goal-directed behavior [Warpefelt and Verhagen, 2017].

Rather than requiring bespoke scripting for every scenario, human-like agents can be seeded into the world with personalities, needs, and objectives, and allowed to act independently [Park et al., 2023]. This creates a more dynamic and persistent sense of place — villagers with daily routines, rival factions with shifting alliances, or lone wanderers with evolving goals. These agents contribute not only to realism, but to systems-driven storytelling that can surprise even the developers themselves.

3.3 Automated play and balance testing

Game balance traditionally requires human playtesters to explore the edge cases and emergent strategies that arise from complex systems. This process is slow, expensive, and often limited in scope. Agents with human-like motivations and varying play styles can serve as surrogate players — automatically exploring a game's mechanics and surfacing imbalances or unintended strategies [Orkin and Roy, 2007].

By tuning the personalities and skill levels of these agents, designers can simulate a wide range of player types: competitive min-maxers, curious explorers, casual roleplayers, and more [Soni and Hingston, 2008]. This allows for more thorough testing across the game's entire design space. Additionally, these agents can serve as long-term regression testers, continuously interacting with the system during development and identifying design regressions or exploits over time [Zhu et al., 2024].

4 Proposed Model Architecture

4.1 The architecture

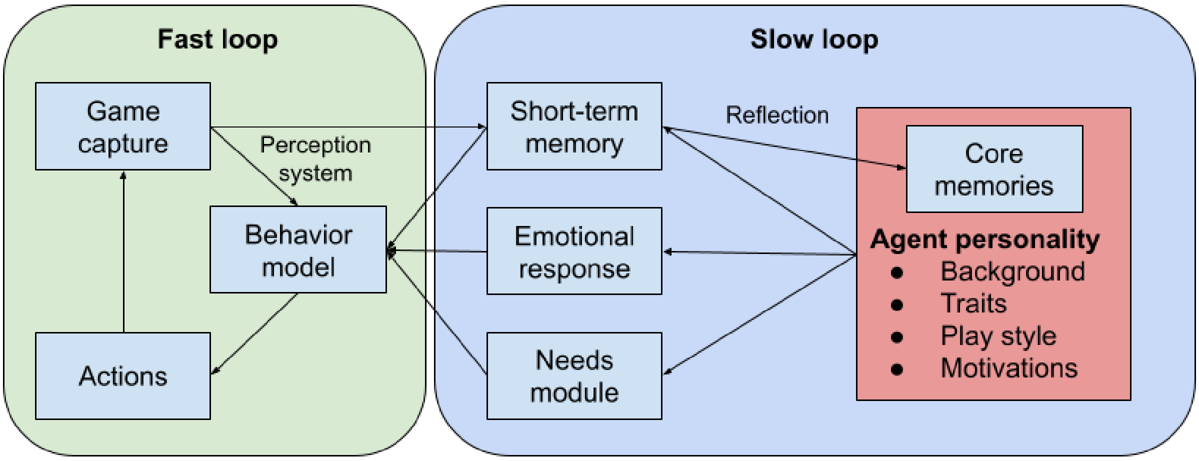

Figure 1: Proposed modular architecture for generating human-like behavior in interactive game agents. The system integrates perception, decision-making, memory, emotion, and agentic traits – such as cultural background, personal history, personality, play style, and skill – into a unified behavioral model.

We propose a modular architecture designed to support the generation of dynamic, context-aware behavior in game agents. This system operates on two primary loops: a **fast loop** for real-time decision-making and a **slow loop** for long-term planning, internal consistency and psychological depth [Rao and Georgeff, 1995, Laird, 2001]. Together, these loops allow agents to react instantaneously to the environment while evolving their behavior based on internal dynamics like personality, needs, and memory.

The architecture consists of the following major components:

- **Perception Layer:** Captures real-time information from the game environment, including sensory inputs and symbolic game state. This layer provides the agent's situational awareness for immediate reactions.
- **Behavior Model:** Translates the internal state representation into concrete actions. This model is responsible for executing decisions in the game world, reflecting the agent's personality and real-time priorities.
- **Emotion Module:** Integrates with the personality module to modulate decisions based on the agent's current emotional state, ensuring that emotions influence behavior in a consistent manner.
- **Memory System:** Divided into short-term memory, which stores recent events and context, and long-term memory, which stores experiences that contribute to personality. Short term memories can become long-term memories through reflection.

- **Personality Module:** Encodes the agent's stable personality traits, ensuring that all decisions align with these core attributes, providing consistency and depth to its behavior.
- **Needs Module:** Models the evolving internal and external motivations of the agent, adjusting over time based on the passage of time and the agent's experiences. This module influences the agent's priorities and decision-making.
- **Long-Term Memory Reflection:** Reflects on the short-term memory and creates core long-term memories that evolve the agent's personality and internal state over time. These memories inform future behavior and strategic decisions.

Information flows through the system as follows: the **fast loop** processes sensory inputs and the game state to execute actions, influenced by the agent's current mental state. The **slow loop** integrates feedback from the environment and approximates the temporal evolution of the agent's mental state, all while being shaped by a core memory or personality definition. The personality definition is also able to evolve over time through reflection from the short-term memory. This creates a system capable of generating both real-time, reactive behavior and deeper, more strategic decisions that evolve as the agent interacts with its environment and other players.

4.2 Data collection

Training and evaluating human-like behavior requires large-scale behavioral data from real players. We propose a hybrid data collection pipeline combining screen capture, structured game state logs, and player input/action traces. Screen captures mimic human perceptual input, while symbolic game state provides structured signals about relevant entities and values.

In addition to moment-to-moment action data, we emphasize the collection of higher-level annotations such as player goals, emotional responses (via chat logs, speech, or behavioral proxies), and strategic context [Orkin and Roy, 2007]. This data can be sourced from playtesters, streamers, or controlled gameplay studies. Longitudinal tracking of players will also allow the construction of datasets reflecting evolving play styles and emergent behavioral patterns over time [Soni and Hingston, 2008].

To address the high cost of large-scale data collection, we propose a complementary strategy based on controlled synthetic generation and inference-driven annotation. By conducting structured gameplay studies with recruited players, we collect rich behavioral traces paired with explicit personality assessments [Xie et al., 2024]. These aligned datasets serve as supervision for training an Inverse Personality Model (IPM)—a predictive model that infers latent personality traits from observed gameplay behavior [Liu et al., 2022]. Once trained, the IPM can be applied to unlabeled gameplay footage from public sources such as YouTube and Twitch, enabling scalable estimation of player personality traits across diverse play styles and environments. In parallel, our agentic framework supports the generation of synthetic gameplay data that reflects consistent internal motivations and emotional variability. These synthetic agents can be labeled via the IPM and distilled into downstream models to improve generalization and robustness across personality-conditioned behaviors.

4.3 Agentic framework

To produce believable and psychologically coherent behavior, we adopt an agentic framework composed of four interacting components: **personality**, **needs**, **emotion**, and **memory**. These components define the internal state of the agent and condition its behavior across time and context [Park et al., 2023].

- **Personality** encodes stable behavioral tendencies that guide how the agent interprets and responds to situations. Rather than building this from scratch, we condition LLM outputs on personality embeddings derived from structured trait models (e.g., Big Five, HEXACO) [McCrae and John, 1992] and gameplay schemas (e.g., Quantic Gamer Motivations).

- **Needs** define the agent's internal drives and priorities. Inspired by frameworks such as the Maslow hierarchy [Maslow, 1943], these needs can be implemented as scalar values or probabilistic activations that modulate goal selection. LLMs can be guided via prompt augmentation or retrieval-based conditioning to reflect needs such as belonging, exploration, or status.

- **Emotion** introduces transient, context-sensitive modulation of behavior. The emotional state is updated based on recent events and tracked as location in valence-arousal space that influences the tone, urgency, or intensity of responses. LLM behavior is adjusted through prompt modifiers or emotion-conditioned decoding strategies to create variability without inconsistency.

- **Memory** enables agents to retain continuity across interactions. Short-term memory captures recent game state, dialogue, and interactions within the current session. Long-term memory encodes evolving beliefs, preferences, and personal experiences [Tulving, 1972]. These memories can be stored as structured state or narrative summaries, which are injected into LLM context windows or retrieved on demand to inform future decisions.

This framework does not require deeply integrated architecture changes, but rather augments the base capabilities of pretrained LLMs with structured state conditioning and lightweight behavioral control. The result is a system that supports consistent, adaptive, and emotionally expressive agents—capable of sustaining the illusion of internal life across extended gameplay.

4.4 Foundation model

As we collect more high-quality gameplay data and refine the agentic control framework, we anticipate the ability to transition from modular orchestration to a fully integrated, end-to-end foundation model. Rather than coordinating behavior through a set of separate components (e.g., emotion, memory, needs), the model will learn to internalize and express these dynamics directly through training.

This foundation model will take as input a structured representation of the game state and a compressed encoding of agent history — including both short-term context and persistent identity traits such as personality and play style [Zhu et al., 2024]. Given sufficient training data, the model will learn to generate contextually appropriate actions and utterances that reflect nuanced, psychologically coherent behavior.

Training will leverage a combination of supervised behavioral cloning, reinforcement learning with human feedback (RLHF), and self-play in both simulated and live multiplayer environments [Bai et al., 2022]. Synthetic data from modular systems and scripted agents can bootstrap learning and accelerate convergence.

The ultimate goal is to produce a behavior engine that captures the expressive range of the agentic framework, but with lower latency, reduced runtime complexity and costs, and greater scalability. As the model architecture improves, we expect it to unify perception, reasoning, and behavioral generation into a single, generalizable agent core capable of producing lifelike behavior across a wide range of gameplay contexts.

5 Model Evaluation

5.1 Embodied Turing Test

Evaluating human-like behavior requires more than task performance metrics—it requires assessing how well the model mimics the emotional, social, and behavioral patterns of real players. To this end, we propose a Turing-style evaluation protocol grounded in social deception: a multiplayer deduction game in which human participants must distinguish between real and AI-controlled players based on behavior alone [Turing, 1950].

In this setting, both human and agent-controlled avatars are placed into a shared game environment. The task of the human players is to identify which entities are AI-agents within a limited time frame, using in-game actions, communication, and movement cues. Agents are trained to behave in ways that are consistent, emotionally reactive, and socially plausible, while humans may issue challenges (e.g., "If you're human, come to the left wall!") to probe for unnatural behavior [Hingston, 2009].

The success of the model is measured by its ability to evade detection — ideally resulting in human accuracy near chance level. This approach allows us to quantitatively evaluate believability across multimodal signals (language, movement, decision-making) while directly optimizing for the illusion of agency. The same protocol can also be used to collect human feedback and fine-tune behavioral outputs via reinforcement learning [Bai et al., 2022].

While evading detection remains the primary success metric, this methodology also enables quantitative evaluation of secondary outcomes such as role alignment, strategic diversity [Xie et al., 2024], and the agents' ability to adaptively shift play styles [Heaton, 1981]. Over time, this evaluation framework will serve as both a benchmarking suite and a source of continual data collection, further improving the foundation model's ability to emulate lifelike behavior at scale.

6 Conclusion

We believe the next major leap in interactive entertainment and AI research lies in building agents that behave not just logically, but **believably** — bots become agents when they feel like they have inner lives, coherent motivations, and the capacity to surprise, amuse, or frustrate us in deeply human ways. Today's NPCs and game bots fall far short of this mark: they are static, mechanical, and emotionally flat. Our framework offers a path forward, one that augments large language models with lightweight, interpretable agentic scaffolding to produce behavior that players can genuinely connect with.

By combining modular simulations of personality, needs, emotion, and memory with the generative power of foundation models, we unlock a powerful synthesis: agents that are both reactive and expressive, both context-sensitive and narratively grounded. This isn't just a UX improvement; it's the foundation for entirely new kinds of gameplay, emergent storytelling, and large-scale simulation [Zhu et al., 2024]. These agents can populate game worlds, test game balance, and evolve alongside human players—not as placeholders, but as participants.

References

[Bai et al., 2022] Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

[Bates, 1994] Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125.

[del Rio-Chanona et al., 2024] del Rio-Chanona, R. M., Pangallo, M., and Hommes, C. (2024). Can generative ai agents behave like humans? evidence from laboratory market experiments. *arXiv preprint arXiv:2505.07457*.

[Heaton, 1981] Heaton, R. K. (1981). *Wisconsin Card Sorting Test Manual*. Psychological Assessment Resources, Odessa, FL.

[Heider and Simmel, 1944] Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57(2):243–259.

[Hingston, 2009] Hingston, P. (2009). A turing test for computer game bots. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(3):169–186.

[Laird, 2001] Laird, J. E. (2001). It knows what you're going to do: Adding anticipation to a quakebot. *AI Magazine*, 22(4):117.

[Liu et al., 2022] Liu, X., Dosovitskiy, A., Brohan, A., Tunyasuvunakool, S., Ghasemipour, S. K. S., Tan, J., Alayrac, J.-B., de Las Casas, D., Vinyals, O., Sermanet, P., et al. (2022). Video pretraining (vpt): Learning to act by watching unlabeled online videos. *arXiv preprint arXiv:2206.11795*.

[Maslow, 1943] Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50(4):370–396.

[Mateas, 1999] Mateas, M. (1999). An oz-centric review of interactive drama and believable agents. In *Artificial Intelligence Today*, pages 297–328. Springer.

[McCrae and John, 1992] McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

[Nass and Moon, 2000] Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

[Orkin and Roy, 2007] Orkin, J. and Roy, D. (2007). The restaurant game: Learning social behavior and language from thousands of players. *Journal of Game Development*, 3(1):39–60.

[Ortony et al., 1988] Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

[Park et al., 2023] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., and Liang, P. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

[Rao and Georgeff, 1995] Rao, A. S. and Georgeff, M. P. (1995). Bdi agents: From theory to practice. In *Proceedings of the First International Conference on Multiagent Systems (ICMAS)*, pages 312–319.

[Smith et al., 2023] Smith, A., Nguyen, T., and Lee, M. (2023). Affective mirroring in video game npcs: A pilot study. *Frontiers in Psychology*. DOI: 10.3389/fpsyg.2023.123456.

[Soni and Hingston, 2008] Soni, B. and Hingston, P. (2008). Bots trained to play like a human are more fun. In *2008 IEEE International Joint Conference on Neural Networks*, pages 363–369. IEEE.

[Tulving, 1972] Tulving, E. (1972). Episodic and semantic memory. In Tulving, E. and Donaldson, W., editors, *Organization of Memory*, pages 381–403. Academic Press.

[Turing, 1950] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

[Warpefelt and Verhagen, 2017] Warpefelt, H. and Verhagen, H. (2017). A model of non-player character believability. *Journal of Gaming & Virtual Worlds*, 9(3):255–272.

[Xie et al., 2024] Xie, Y., Li, Z., Wang, X., et al. (2024). Be.fm: Open foundation models for human behavior. *arXiv preprint arXiv:2505.23058*.

[Zhu et al., 2024] Zhu, R., Huang, Y., Peng, B., et al. (2024). Large language models and games: A survey and roadmap. *arXiv preprint arXiv:2403.12345*.