The Many Challenges of Human-Like Agents in Virtual Game Environments

Maciej Świechowski
QED Software
Warsaw, Poland
maciej.swiechowski@qed.pl

Dominik Ślęzak
University of Warsaw
Warsaw, Poland
slezak@mimuw.edu.pl

ABSTRACT

Human-like agents are an increasingly important topic in games and beyond. Believable non-player characters enhance the gaming experience by improving immersion and providing entertainment. They also offer players the opportunity to engage with AI entities that can function as opponents, teachers, or cooperating partners. Additionally, in games where bots are prohibited – and even more so in non-game environments – there is a need for methods capable of identifying whether digital interactions occur with bots or humans. This leads to two fundamental research questions: (1) how to model and implement human-like AI, and (2) how to measure its degree of human likeness.

This article offers two contributions. The first one is a survey of the most significant challenges in implementing human-like AI in games (or any virtual environment featuring simulated agents, although this article specifically focuses on games). Thirteen such challenges, both conceptual and technical, are discussed in detail.

The second is an empirical study performed in a tactical video game that addresses the research question: "Is it possible to distinguish human players from bots (AI agents) based on empirical data?" A machine-learning approach using a custom deep recurrent convolutional neural network is presented.

We hypothesize that the more challenging it is to create human-like AI for a given game, the easier it becomes to develop a method for distinguishing humans from AI-driven players.

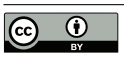
KEYWORDS

Intelligent Agents; Human-Like AI; Believable Agents; ML Model for Turing Test

ACM Reference Format:
Maciej Świechowski and Dominik Ślęzak. 2025. The Many Challenges of Human-Like Agents in Virtual Game Environments. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

ACKNOWLEDGMENTS

This research was co-funded by the Smart Growth Operational Programme 2014-2020, financed by the European Regional Development Fund under a GameINN project POIR.01.02.00-00-0207/20, operated by The National Centre for Research and Development.



Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Games have been an integral part of human civilization since ancient times [23]. In the essay "Homo Ludens", the author examines games as a fundamental condition for the evolution of culture [11]. Besides providing entertainment, some games help in training the mind, enhancing eye-brain coordination, and improving reflexes.

Since the advent of Artificial Intelligence (AI), games have served as a testbed for its development. Initially, AI research focused primarily on *chess* [31] and *checkers* [41]. However, with recent advancements that have sparked debates about the human-likeness of AI, video games have become increasingly attractive as research environments. Unlike abstract combinatorial games such as *chess*, *checkers*, or *Go* [14], video games typically offer a simplified model of the real world populated with numerous non-player characters (NPCs). They were named "*human-level AI's killer application*" [25].

The aim of this article is twofold and is divided into two parts. The first part, contained in Section 2, provides a comprehensive and accessible review of the challenges related to creating human-like AI (characters, bots, players) for games. We discuss 13 challenges, drawing from both the literature and our years of professional experience in implementing artificial intelligence for video games. These challenges can be generalized to autonomous robots and any virtual environments with intelligent agents. We believe that the comments and insights provided mostly in this part can help researchers design AI that acts more human-like by addressing the issues related to creation and evaluation. In our search for relevant papers, we queried popular bibliographic databases using terms from the following template: ⟨ adjective ⟩ ⟨ noun ⟩. The set of adjectives included "human-like", "human-level", "believable", and the set of nouns contained: "agent(s)", "player(s)", "bot(s)", "character(s)", "behavior", "AI", and "Artificial Intelligence". The complete set of terms was a Cartesian product of the sets of adjectives and nouns. Additionally, we included other selected articles such as [44] that are seminal to the topic. Among all the papers matching the query, we read their abstracts and introductions to verify if their contents were related to creating or evaluating human-like AI in virtual environments. After this step, 54 papers remained. We have analyzed them and distilled the most pertinent and common challenges.

Parallel to creating human-like autonomous agents, we also discuss the issue of assessing their human-likeness. These topics are closely intertwined. It is challenging to implement AI techniques without precisely setting the evaluation criteria—what constitutes human-likeness. In 1950, Alan Turing posed the question "Can Machines Think?" [55]. He introduced the concept of the *Imitation Game*, which laid the foundations for what would later be known as the *Turing Test* [35]. In the classic variant of the test, an interrogator (judge) interacts with two players, *A* and *B*, using a natural

language chat interface in such a way that the players cannot be seen—only their responses can. The goal of the interrogator is to determine whether each participant is a human or a computer. The *Turing Test* has been a foundational concept in AI and has sparked a debate about whether machines can display human-like abilities, intelligence, and consciousness. This topic has been pursued by many prominent researchers, such as Lofti Zadeh [58]. Although there have been attempts to formalize believability [4], the most common approach is to propose a *Turing Test* analogy for video games [47]. The first research framework to do so was the *2K BotPrize* [18], proposed in 2008 by Philip Hingston, based on the multi-player shooter game *Unreal Tournament 2004*. Since then, researchers have adopted the idea of the *Turing Test* for more games, such as *Street Fighter* [2] and *Ms. Pacman* [34].

In the second part, we present a study concerning the automatic construction of a method capable of distinguishing human players from bots solely by learning from data. In Section 3, the environment in which the experiments were conducted is presented—a tactical war video game with relatively high action-space complexity. Section 4 focuses on the machine learning algorithm proposed for this task. The training methodology combines recurrent and convolutional neural networks and utilizes multi-modal (numeric and spatial) data. The proposed solution achieves an F1-Score of 0.92, which is a significant improvement over the previous solution based on XGBoost with only numeric features, which had an F1-score of 0.58. The last section is devoted to conclusions.

2 DISCUSSION OF THE CHALLENGES

Humans are Diverse. The first challenge we wish to highlight is the ambiguity and imprecision inherent in defining the goal of human-like AI. Following [2], human-like AI is described as "behaving in a manner that makes it indistinguishable from human players". The authors of [50] define it as "giving the feeling of being controlled by a (human) player". Considering the vast diversity among humans, what exactly does it mean to play like a human player? The playing styles can vary significantly. An elderly individual often plays differently compared to a younger person or a child who just started playing video games. Similarly, a professional player's approach will differ from that of an amateur. Even among seasoned players, there is a lot of variety. The authors in [1] identify seven distinct player types: *power gamer*, *butt-kicker*, *tactician*, *specialist*, *method actor*, *storyteller*, and *casual gamer*. Another example is the Bartle taxonomy, outlined in [59], that categorizes players into four roles: *killer*, *socializer*, *achiever*, and *explorer*. Each category adjusts their gameplay in a unique manner to align with their personal objectives and maximize enjoyment.

Unreal Tournament 2004 serves as a popular choice among researchers for assessing human-like behavior in gameplay. According to [13], one of the significant challenges for human judges is the wide range of human behaviors, which underscores the complexity due to player diversity. The authors investigated the topic further and commented that skill level is the most crucial factor distinguishing different playing patterns. In their study, less skilled players were identified as outliers in a statistical analysis of gameplay when matched with experts.

Additionally, implementing AI bots to mimic human behavior can be problematic if the game is still under development, because it might be unclear how human players will play the particular game. This becomes particularly challenging for novel games, i.e., not based on standard repeatable formats.

The Complexity and Expressiveness of Action Space. In short, complex high-dimensional action spaces pose a great challenge in implementing human-like AI, whereas simple, constrained ones complicate the evaluation of whether an AI is truly human-like. Let us elaborate on this.

In complex environments, human-like behavior necessitates the simulation of thinking processes. These usually involve various forms of reasoning and strategic or tactical planning that anticipate multiple steps ahead. Humans often develop behavior patterns based on their experiences and can rely on their intuition. In contrast, AI-driven characters depend on computational techniques such as tree search (e.g., alpha-beta or Monte Carlo Tree Search (MCTS) [48]), planning (such as Hierarchical Task Networks (HTN) [37]), or machine learning. The complexity of the environment significantly increases the computational resources required to create even a moderately proficient AI player compared to human players. This is clearly demonstrated by historical attempts to challenge top human players in various games, such as *Chinook* [42] (in checkers), *Deep Blue* [7] (in chess), *AlphaGo* [45] (in Go), *AlphaStar* [52] (in Starcraft II), and *OpenAI Five* [38] (in Dota 2). These projects employed substantial computational resources. For instance, Deep Blue was ranked 259th on the top500 list of supercomputers, while OpenAI Five utilized 172,800 CPUs and 1536 GPUs simultaneously [38] at a peak. Although creating a potent computer agent is not equivalent to developing human-like AI, these research projects began their efforts when the top programs were significantly inferior to the skills of amateur human players. In complex games, achieving a skill level comparable to an average human player also demands considerable computing power and/or extensive simulation times. The complexity originates from several aspects: the number of options available in each state (termed the branching factor), the length of the game (which necessitates longer simulations, thereby reducing the number that can be conducted within a given time), the diversity of game states (state-space complexity), and the total number of leaf nodes in a complete game tree (game-tree complexity). Chess, being the second least complex game in this context (after checkers), possesses an average branching factor of 35, a state-space complexity of 10^{44}, and a game-tree complexity of 10^{123}. To provide a sense of scale, it is estimated that there are 10^{80} atoms in the observable universe. The complexity escalates further in real-time games with continuous movement making the action space virtually infinite (as bots can go in any directions). In the study concerning believable navigation [21], the authors found that AI players can get stuck on level geometry or fail to appear human when following the navigation graph. The authors of [2] emphasize the necessity of exploring high-dimensional action spaces as one of the two main challenges in creating human-like AI.

The five aforementioned AI players were the outcomes of large-scale research projects focused on competing against top human players in dedicated matches. However, AI players integrated into

commercial video games must operate within the constraints of a single consumer-level hardware system shared with other game functionalities such as rendering [39]. This restriction makes it particularly challenging to develop AI players equalling humans in skills without giving them an unfair advantage.

Now, let us focus on simple abstract game environments characterized by low complexities in both state space and action space. *Tic-Tac-Toe* serves as a good example. While it is straightforward to create competent or even exceptionally strong AI players for such games, differentiating whether a player is human-like proves challenging. If there are only a handful of actions and the game follow strict rules, e.g., players take turns and each turn the active player chooses one of the few available actions, then there are insufficient premises to distinguish humans from bots. This issue is even more pronounced in *Rock-Paper-Scissors*, where each action is equally viable, making even random choices a legitimate strategy. In the study [34], the authors highlighted the difficulties in telling whether a player is human or a bot within the game *Ms. Pac-Man*, which is considerably more complex than *Tic-Tac-Toe*.

The underlying point is that a game or virtual environment must possess a sufficient level of expressiveness to enable feasible assessments of human-like behavior. The more open-ended an environment is, the easier it becomes to determine whether the behavior of agents within it is believable.

The Challenges of Scale. As mentioned in the previous paragraph, creating human-like AI is extremely computationally demanding. Note that all examples of AI agents that achieved high efficacy were in 1-vs-1 player settings. In these cases, the AI, utilizing powerful hardware, controlled one player (or, at most, two during training via self-play).

Now, imagine a virtual city populated by a million NPC inhabitants, each individually controlled. These NPCs must take actions based on various circumstances such as their goals and both internal and external states. One of the models for simulating many agents is called Belief–Desire–Intention (BDI) [15]. Scaling high-efficacy AI implementations to environments with many diverse agents presents a challenge that is multiple orders of magnitude greater [3]. Essentially, it requires multiplying the resources used to create a high-fidelity solution by the number of simulated agents. As of the time of this article, it is infeasible to apply *state-of-the-art* game-playing models to massively multiplayer games. Currently, dedicated, low-fidelity approaches known as "Crowd AI" are used in the industry. Nevertheless, there have been research attempts to tackle the problem of simulating a large number of believable bots, as discussed by [40]. This approach utilizes Utility AI, integrates the Observe-Orient-Decide-Act (OODA) decision cycle, and employs temporary roles that agents may assume.

Avoiding Superhuman Behavior. Rational players must act towards their goals and posses a certain level of skill in order to be believable. In efforts to develop AI agents with sufficient competence, there is a risk that these agents may display superhuman abilities in certain aspects of a game, compromising their believability. Both [8] and [28] highlight that precision in calculations, such as pixel-perfect aiming in shooter games, is a characteristic strongly associated with bots in scenarios analogous to the *Turing Test* for video games. Therefore, it is another feature that makes it easier to assess human-like behavior but more difficult to implement it. It is often the case, that the removal of such precise calculations from bots make them significantly weaker, to the point that they are not believable due to different reasons, i.e., their incapacity to perform competent actions.

Idle and Non-Relevant Actions. Even when a game has a clearly defined objective, human players often engage in actions that are irrelevant from the perspective of this goal. Such behaviors are commonly described in the literature as roaming, idle, "for fun", or for "own amusement" actions. Observing navigation in an open virtual world provides a particularly interesting context to study these patterns [32]. For example, players exploring a city in an RPG might pause to admire specific 3D models within the game environment. The following example, based on the author's experience in a *Diablo* game developed by *Blizzard Entertainment*, illustrates this point. Imagine a scenario where a player controls a character equipped with an artifact that leaves a trail of ice on the ground. Players might intentionally run in patterns to create specific shapes with this mechanic. This emergent behavior, stemming from human creativity and playfulness, is entirely disconnected from the game's mechanics or objectives. Consequently, it presents significant challenges for implementing such behaviors in AI-controlled bots. Standard algorithms designed to enhance AI player proficiency, which is typically the primary focus, would likely treat these actions as noise – something irrelevant. The study mentioned in [22] proposes a modification to the Monte Carlo Tree Search (MCTS) algorithm that biases action selection towards patterns observed in human gameplay to make it more human-like. Given that enough data is available from human players, this adaptation appears to be a promising strategy for developing more human-like AI behaviors.

Biological Constraints. Equipping AI players with realistic biological constraints extends well beyond standard practices in the field [33, 39]. These constraints include:

- *Visual Perception.* Typically, AI-controlled players are fed information directly about game objects and other characters, without any simulation of actual perception. For example, one of the distinguishing factors between bots and humans in UT2004 is that bots often collect items not looking at them [43]. The article [24] discusses the challenge of extracting spatial information about the physical environment – such as walls and doors – from the game's internal data structures, which are just sets of polygons.
  As an interesting side note, the authors of [51] provided an insightful finding in their paper. The assessment of a bot's human-likeness, as judged by humans, showed significant variation depending on the observation perspective. The authors concluded that a more accurate assessment occurred when bots were observed from a third-person perspective, as opposed to a first-person perspective.
- *Sound Perception.* Simulating realistic sound perception is as challenging as visual perception. The detection of sounds by a bot should not be binary, it should involve some level of

fuzziness and imperfection, similar to the senses of living organisms. For instance, the bot should be able to express uncertainty by saying, "I didn't hear you."

- *Memory*. In the article "Do Non-Player Characters Dream of Electric Sheep?" [20], the author emphasizes memory functions as a crucial component for creating believable NPCs. Designing a memory system that encompasses prioritization of information, realistic forgetting, and loss of detail while retaining key facts presents significant challenges. These aspects must be managed without undermining the goal-oriented performance of AI players, a recurring theme in this section.
- *Cognitive Load and Ability to Multitask*. In real-time video games, especially strategic ones, AI players often display less strategic reasoning than humans. However, they compensate by being able to oversee the entire map and control numerous units at once, which would be overwhelming and physically impossible for human players. Simulating realistic constraints related to attention focus and multitasking poses multiple challenges: determining appropriate game-specific limits, deciding which aspects of the game the AI should focus on, and the fact that imposing such constraints might diminish the AI's effectiveness. Creating competent bots that operate on consumer-level hardware remains a significant challenge.
- *Reaction Time*. Reaction time serves as another biological constraint. In video games, AI players frequently benefit from seemingly unlimited reaction times, a key feature distinguishing bots from human players [29]. There have been propositions to limit the reaction time and action rate of AI players, as seen in [52].
- *Other Constraints*. The article titled "Video Game Agents with Human-like Behavior using the Deep Q-Network and Biological Constraints" [36] introduces additional constraints such as "confusion" (experienced when suddenly surrounded by many enemies), "fluctuation" (errors in operation), "delay" (akin to reaction time), and "tiredness".

Emotional Element. The authors of [12] wrote:

> "*Problems of NPCs usually lie in their lack of convincing social and emotional behaviour raising the need for a robust affect module within the agent's architecture. Developing an integrated architecture would ideally require developing models for the theory of emotion, social relation, and behaviour, and combining the theories into an overall model.*"

In [26], emotions are listed as one of the most critical qualities of believable bots, ranking just after goals. The authors of [5] identify emotions as one of the seven believability characteristics, alongside personality, self-motivation, change, social relationships, consistency of expression, and illusion of life.

Emotions have been extensively studied in psychological research [19]. However, there are currently no established computational models to accurately simulate emotions in video games, making this issue particularly challenging. There are many facets to emotions in games. An emotional response may lead to changes in a player's behavior following specific in-game events. The next example will be from the game *Tactical Troops: Anthracite Shift*, discussed in a study in the next section of this article. In this game, a player wins after eliminating all enemy units. We observed a scenario where an AI-controlled player had two units remaining, and the opposing player had only one. One of the AI's units had a clear shot at the enemy but was blocked by another friendly unit. The AI chose to fire anyway, eliminating both the friendly and the enemy unit, thus securing a win. This behavior, while effective, is rarely observed in games played by humans, who typically avoid harming their own units.

Another aspect of emotional display in gaming is the anger or "tilt" effect [57], which refers to a state of emotional frustration or upset that negatively impacts a player's performance. This phenomenon occurs when players become agitated due to a series of losses, perceived unfairness, or other in-game setbacks, leading to increasingly poor decision-making and potentially aggressive behavior. The term "tilt" originates from poker but has become widely used across various competitive gaming genres.

Handling Uncertainty. Uncertainty in games typically involves hidden information (asymmetric information between the players), randomness, or both. This already poses significant challenges for creating agents aimed at achieving effective play because it results in a combinatorial explosion of potential game states. In research related to game AI, this issue is generally addressed using determinization techniques [10] and by replacing perfect information game states with information sets [9]. In the video game industry, the amount of hidden information is often so vast that AI characters are usually given unfair access to it, albeit with some techniques implemented to make this less apparent [27]. For example, in real-time strategy (RTS) games, most of the map is concealed from the players by the so-called "fog of war." Players can send units to anywhere on the map to scout and reveal the covered areas. AI players, however, are typically given hints about a few potential locations of the human player's base, including the correct one, thereby reducing the amount of scouting required. The same applies to information about the resources and military capabilities the human player possesses.

Uncertainty introduces an additional layer of challenge when designing human-like bots. If bots perform too efficiently, for instance, by employing complex probability estimations, they might be perceived by human players as cheating AI [24], even if this is not the case. To counteract this, a solution is to integrate a human-like reasoning process for inferring hidden information [30]. However, this approach is computationally intensive and requires significant trade-offs in terms of playing efficacy.

Adaptability. The article [28] hypothesizes that AI in commercial games is often exploitable due to its tendency to follow repetitive patterns, thereby allowing human players to recognize and adapt to these behaviors. Meanwhile, the authors of [6] assert that "adding some unpredictability can significantly enhance believability," and the authors of [46] demonstrate a strong correlation between unpredictability and human-like behavior.

In [2], the authors identify "adaptation" to the opponent's style as one of the two primary challenges in developing human-like AI for the *Street Fighter IV* game. The other challenge they outline is

exploring a high-dimensional state-action space. To address this, they implemented real-time reward transformations within their reinforcement learning framework. This strategy involved dynamically altering the reward definitions based on the performance of various playing styles. As a result, they achieved a human-likeness score of 0.64 on a scale from 0 to 1, slightly less than the top score of 0.67 achieved by a human player, significantly surpassing the baseline bots, which scored between 0.28 and 0.30.

Making Mistakes. The analysis of literature related to the creation and evaluation of human-like AI underscores that human players indeed make mistakes. As noted by the authors of [12]:

> "*An intelligent agent model should not require producing a "perfect" agent, but rather, for better human resemblance and higher believability, it is more natural to have the flaws and dysfunctionalities of the human affect phenomena incorporated into the model.*"

However, humans learn, and there is significantly less likelihood of repeating the same mistakes. In the *UT2004* competition, bots that repeated the same mistakes consistently were quickly judged as non-human-like [43].

Implementing both the inclusion of mistakes and learning from them is challenging. Mistakes can emerge unintentionally without being explicitly programmed into bots, but these unintentional errors can be challenging to identify and incorporate with adaptation mechanisms. In addition, they can be of artificial nature such as bots getting stuck in level geometry [43], which was a telling factor for judges responsible for the video game *Turing Test* in *UT2004*. Deciding which aspects of gameplay should intentionally involve mistakes in a believable manner remains a significant challenge.

Training Human-Like AI. One of the most promising approaches to creating human-like AI agents is training them using machine learning methods (ML). When the objective is to develop the strongest agent possible, training can be conducted using reinforcement learning without human knowledge [45, 53]. In this setup, agents compete against different versions of themselves, continually refining their skills. However, when aiming for human-likeness or believability, training – whether supervised or through reinforcement – using human games becomes a more suitable approach [2].

Training believable agents from human data presents many of the already mentioned challenges:

- *Humans are diverse*. Training can utilize games played by either a specific group of players or the general population. In the latter case, the resultant behavior will likely be an average.
- *High computational demand*. Vast action spaces, especially in reinforcement learning, are known for their sample inefficiency and consequent computational expense [17].
- *The effect of scale*. The trained model must be inferred as often as there are intelligent agents in the environment. However, modern GPU-based ML models facilitate batch processing.

Additional challenges specific to training include:

- *Large volumes of training data are required*. This data is usually not available for games at the time of development. As the game has not yet been released, the only data available typically comes from developers and testers playing it, which is insufficient for large-scale training.
- *Human-Like Control Input*. In video game development, AI systems typically process inputs differently from human players, which poses a fundamental challenge in creating believable behaviors. Humans use controllers such as: keyboard, mouse, game-pads (including analog joysticks), steering wheels, etc. Using a controller involves atomic actions, e.g., pressing a key or applying a force in one or more of the controller axes. The observable in-game behaviors emerge from sequences of these inputs, adding complexity to the input space. In contrast, AI in games is usually programmed to perform high-level behaviors like "move to point X", "attack enemy Y", or "flee", which operate on a more abstract level than direct input actions.

Simulating Social Norms. In the literature related to human-like AI agents, the simulation of social norms is frequently highlighted as an important factor for enhancing believability [20, 54, 56]. In [20], the term "socially believable" is used to describe agents that effectively cooperate and coordinate within a group. The authors in [56] enumerate several social cues essential for believable agents, including: reflexivity, grouping, attachment, reciprocity, and the ability to attribute mental (intentional) states to oneself and to others.

In [54], the concept of a *Game Agent Matrix* is introduced. This matrix includes columns labeled *single agent*, *multiple agents*, *social structural*, *social goals*, and *cultural historical*, and rows labeled *act*, *react*, and *interact*. The matrix cells incorporate concepts such as *awareness*. The study evaluated many games, finding that numerous concepts, particularly those related to social structural and social goals, were underrepresented. The authors emphasized:

> "*NPCs need to exhibit behavior consistent with the environment, the situation and their character in order to seem believable.*"

Incorporating social norms into video games presents substantial challenges. A notable example is *The Elder Scrolls IV: Oblivion*, developed by *Bethesda Software*. In this game, despite the social norm against stealing, players can exploit a loophole by placing a basket over a shopkeeper's head, preventing them from noticing thefts. This is an interesting attempt at introducing realistic perception mechanics, yet it lacks a robust implementation of social norms.

Specific Human-Like Activities in the Game. Let us conclude this section with the observation that games (or virtual environments in general) are highly diverse and open-ended. They may include any aspect of real-world activities, such as conversing in natural language or driving. All such in-game activities can be assessed for their human-likeness [16]. On one hand, this diversity makes the task of creating general solutions for testing AI believability very challenging. An AI-based *Turing Test* judge should either be customized for a specific activity or capable of evaluating every task humans perform. On the other hand, these specific in-game activities could make it easier for human judges to distinguish between humans and bots. Even if bots are competent and believable in core gameplay, they might fail at some specific tasks, such as engaging in natural conversation or driving on streets.

3 EXPERIMENT ENVIRONMENT

The rest of this article is dedicated to a study concerning the problem of believable bots – their distinction from human players and evaluation of their human-likeness ratio in a team-based tactical commandos-style game. The game chosen for this study is *Tactical Troops: Anthracite Shift* [49] (depicted in Fig. 1), which is available on the *Steam* platform[1].

In *Tactical Troops: Anthracite Shift*, each game is played on a 2D map viewed from above. Two players alternate turns, each commanding up to four units. Units are eliminated when they lose all their health points (HP). Unlike many turn-based games, movement is continuous rather than grid-based. The maximum distance a unit can move is determined by its action points (AP), which are also used for performing actions. For example, the larger circle in Fig. 1 illustrates the maximum movement range in a single turn, whereas the smaller (inner) circle indicates the range within which a unit can move and still fire its current weapon.

Figure 1: Movement range in Tactical Troops: Anthracite Shift.

Units can perform actions in any sequence. Considering the main focus of this article, which is the assessment of human-likeness, it is essential to discuss the potential actions each unit can undertake:

(1) **Movement**: Effective movement is crucial as the map's dynamic environments feature buildings for cover, strategic hiding spots, explosive elements, and teleporters that allow for rapid relocation between a designated pair of positions.
(2) **Shooting**: Each unit is equipped with two weapons from a selection of over 30 types. The shooting action includes choosing a weapon, selecting a shooting mode (single or burst), and positioning accurately (units shoot in the direction they face).
(3) **Reloading** weapons.
(4) **Using gadgets**: Units carry up to three gadgets, which can be throwable (e.g., grenades, mines) or togglable (e.g., cloaks, shields, and armors). Using throwable gadgets effectively requires careful consideration of positioning and the force applied.
(5) **Overwatch**: This stance transforms a unit into a stationary defense that automatically fires at the first enemy unit entering its range during the opponent's turn.

The game features two alternative victory conditions:

(1) **Elimination**: Defeat all enemy units.
(2) **Domination** or **Devastation**: In Domination, a player wins by controlling the majority of designated areas (control points) on the map for an entire turn. Usually, a map contains three control points, and controlling at least two is necessary for victory. Devastation involves the destruction of specific stationary objects on the map. While the Elimination condition is always applicable, Domination and Devastation modes are exclusive to specific maps. Introducing these modes aims to discourage defensive play ("camping") and promote more dynamic interactions.

3.1 Method Behind AI Agents

For a comprehensive explanation of the AI player's methodology in *Tactical Troops: Anthracite Shift*, please refer to [49]. This method involves a hybrid approach combining Utility AI [39] and Monte Carlo Tree Search (MCTS) [48], which are integrated through a blackboard architecture. The Utility AI component handles the strategic layer at a high level. It evaluates and assigns one of six possible orders to each controlled unit, where some orders may include parameters like assaulting or defending a specific point of interest. The utility value of each order is determined by a real-valued numerical score, derived from 10 different considerations (factors). Orders are assigned to units on a greedy basis, meaning that the first unit to receive an order is the one with the globally maximum among each unit's highest scored orders. Assigning an order triggers a recalculation of scores for all other units.

The MCTS algorithm acts as the tactical layer, employing a simplified model of the game. The execution quality of a strategic order is used as a heuristic evaluation function in MCTS, which allows for early termination of simulations. To manage the complexity and avoid combinatorial explosion, the sequence in which units are simulated by MCTS is heuristically determined at each turn. MCTS is allocated a budget of approximately 30, 000 iterations per turn.

4 METHOD BEHIND HUMAN-LIKENESS EVALUATION

In an initial study [49], an XGBoost model utilizing 20 summary features per match was trained to differentiate human players from bots. It achieved an accuracy of 0.68 and an F1-score of 0.58. This model was developed using 800 matches and evaluated on an independent set of 200 matches.

In this study, we present a more sophisticated approach using a hybrid deep neural network that combines convolutional (CNN) and recurrent (RNN) subnetworks. The purpose of the CNN component is to extract features from 2D images, while the RNN component is dedicated to processing summarized data and detecting dependencies within it. The full architecture of the solution is presented in Fig. 2.

[1] https://store.steampowered.com/

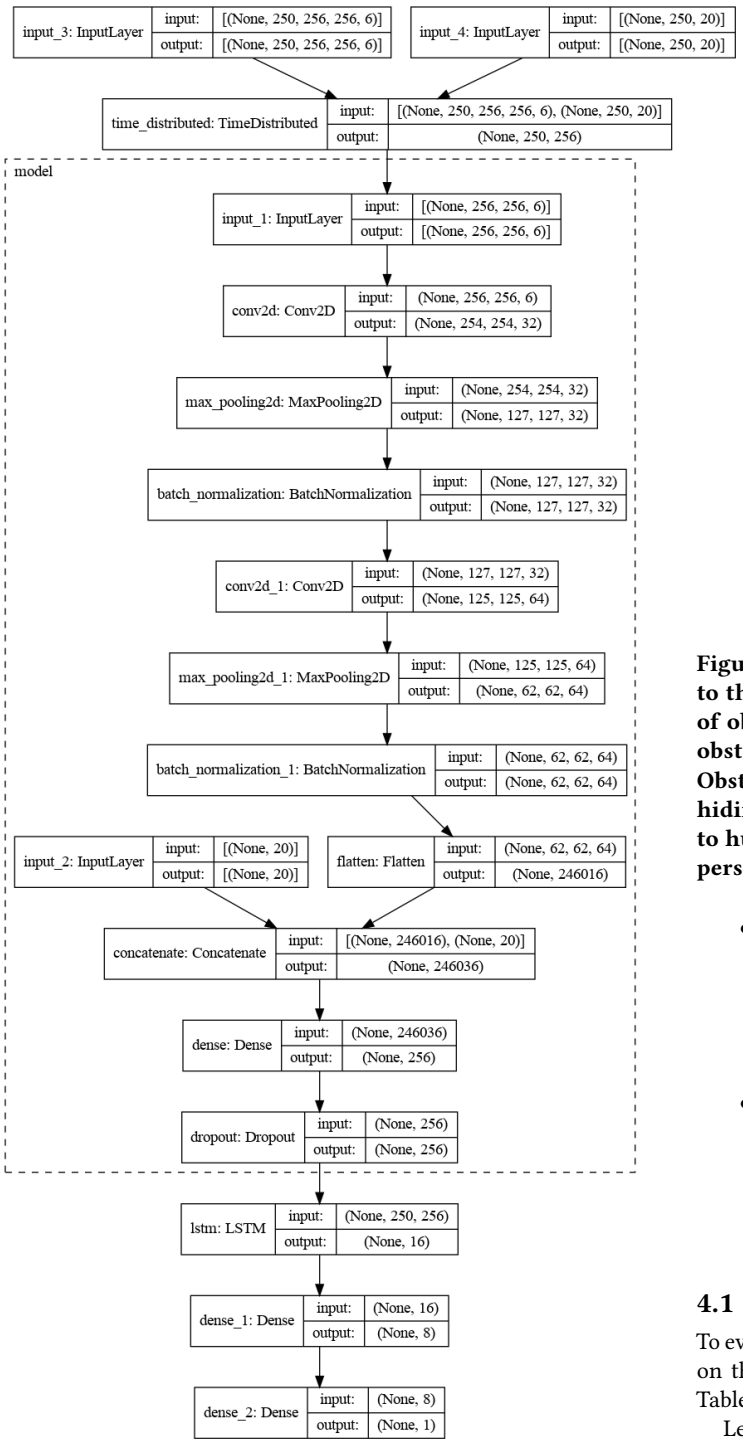


Figure 2: Neural Network architecture for the problem.

We will now give an overview of the input.

- *Input_1*: graphical (pixel) representation of the map. It involves 6 different layers (submaps): obstacles (see Fig. 3), rooftops, teleports (that influence movement capabilities), control points, friendly units with health (the health values are normalized and represented as a color saturation), enemy units with health values. The last two maps are generated dynamically, based on the current state of the game (in the current time).

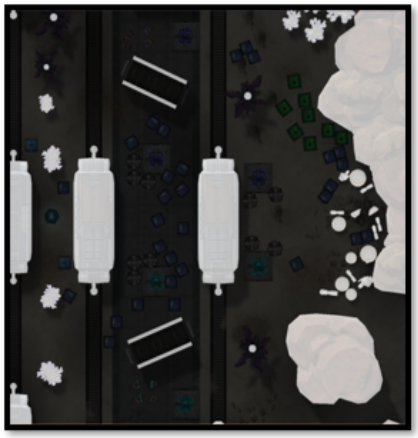


Figure 3: A map of obstacles. One of the six input layers to the spatial component of the neural network. The map of obstacles is crucial because the shapes and positions of obstacles affect lines of sight and lines of fire for the units. Obstacles play a key role in unit placement such as finding hiding spots and avoiding exposure or finding good places to hunt other units. Therefore, they are essential from the perspective of making intelligent positioning decisions.

- *Input_2*: vector representation of 10 x 2 (for both players) numerical features that compute certain values until the current time t. These values are: turn number, damage dealt and received, friendly-fire damage, friendly-fire to total damage ratio, # of used grenades, damage dealt using grenades to total damage, # of used gadgets, # of units' status changes.
- *Input_3 and Input_4*: recurrent neural layers for the spatial and numerical representations, respectively. To properly capture the dynamics of game-play, the input data is passed to the network repeatedly as a sequence of the last 250 states of the game. If a particular game is shorter, then the missing inputs are replaced by zeroes.

4.1 Results

To evaluate the efficacy of the model, we used 5-fold cross validation on the set of available game logs. The results are presented in Table 1.

Let the possible classes be $\{Human, Bot\}$. Let TP_{Human} denote the number of true positives calculated for the human class. Analogously, TN_X, TN_X, TN_X denote true negatives, false positives, and false negatives. In addition to Macro-F1 score defined as:

$$Macro|F1 = \frac{F1_{Human} + F1_{Bot}}{2} \tag{1}$$

where:

$$F1_X = \frac{2 * TP_X}{2 * TP_X + FP_X + FN_X} \tag{2}$$

we also calculated precision for the human class:

$$Precision(Human) = \frac{TP_{Human}}{TP_{Human} + FP_{Human}} \quad (3)$$

and recall for the human class:

$$Recall(Human) = \frac{TP_{Human}}{TP_{Human} + FN_{Human}} \quad (4)$$

The training data for the game were imbalanced due to the limited availability of human players. It comprised a total of 93,195 logs from 89,667 AI vs. AI matches, 2,190 AI vs. Human matches, and 1,338 Human vs. Human matches. Due to this imbalance, per-class metrics for the human class are included, as this is a minority class and it is more susceptible to higher errors.

Table 1: Results achieved using 5-fold cross validation.

Model	F1 Score	Precision (H)	Recall (H)
Full (RNN+CNN)	0.92	0.87	0.81
RNN only	0.88	0.75	0.79
CNN only	0.59	0.61	0.57

The proposed method, that combines CNN and RNN architectures, achieved an F1-Score equal to 0.92. This result clearly indicates that the deep learning model significantly outperformed the initial XGBoost model, which was based on only 20 input features, by improving the F1-Score from 0.58 to 0.92.

In Table 1, there is also a comparison of the full approach with models based on a single component only: either the RNN network transforming aggregated scalar features from the game states or the CNN network with six types of 2D maps as input per game state. We can observe that although a considerable amount of information can be inferred from the maps (e.g., the damage dealt thanks to unit health heatmaps or the presence of units at a particular game state), the CNN model performs significantly worse for the task, achieving an F1-score of only 0.59. The RNN network alone is slightly inferior to the full approach in terms of F1-score but exhibits significantly lower precision. Combining the RNN and CNN components drastically enhances precision and slightly improves the F1-score as well.

Our method achieves better results in differentiating human players from bots compared to most results reported in the literature for different games. For instance, in Pac-Man [34], the differentiation had a success rate of 74%. However, it is inappropriate to compare human-like assessment models for agents across different games. As discussed in Section 2, not only do different environments allow for different expressions of believability, but also the perception of it depends on how the AI was created. For example, both extremely weak and extremely strong AI agents would likely be considered not human-like. The conclusions stemming from these observations will be discussed in the next section.

5 CONCLUSIONS

Games have transcended mere entertainment. They also serve as experimental platforms for various algorithms and AI techniques. Developing models to distinguish accurately between humans and AI bots presents unique opportunities in several domains including fake information detection, identity verification (e.g., in educational, healthcare, financial transactions, creative industries, job recruitment, and online dating), scam prevention, and other illegal or immoral activities, along with measuring AI progress. The inspirations trace back to the *Imitation Game*, potentially paving the way for next-generation *CAPTCHAS*.

In the first part of the paper, we explored various aspects involved in the creation and evaluation of believable AI agents. These include the diversity of human players, the complexity of the action spaces, the challenges of scalability, avoiding superhuman capabilities, introducing idle and non-relevant actions, considering biological constraints, incorporating emotional elements, handling uncertainty, adaptability, allowing bots to make and recover from mistakes, training human-like AI, simulating social norms, and challenges related to specific human-like activities within the game.

In the subsequent part, we addressed the challenge of constructing a model for assessing AI in terms of human-likeness within the game *Tactical Troops: Anthracite Shift*. By integrating deep convolutional and recurrent neural networks, we achieved an F1-Score of 0.92, marking a substantial enhancement from a prior approach that utilized XGBoost and a simpler training configuration. The proposed architecture is relatively general (combining visual spatial data with numeric data), so it can serve as an inspiration for ML models for similar tasks.

We propose an open hypothesis that **the complexity involved in creating human-like agents in a particular environment correlates inversely with the ease of developing methods to distinguish between humans and bots in that environment**. We invite researchers to engage with this hypothesis to validate or refute it.

We contend that a human-likeness assessment model can serve as an automated quality assurance tool in the development of AI players. This model and the agents can be iteratively refined. The process would start with creating weak bots, which the model would not classify as human-like, and progressively aim to deceive the model through enhancements. Once the model erroneously classifies these improved bots as humans (false positive), it would be judicious to retrain the model.

The presence of human-like NPCs in games offers multiple benefits. Primarily, they enhance the gaming experience by balancing realism and playing strength [46]. Additionally, they can reduce the "cold start" effect in massively multiplayer online games until a critical mass of human players is achieved. Furthermore, they facilitate more precise testing by simulating human player behaviors. Human-likeness assessment models also have broader applications, such as bot detection in environments where bot usage by players is considered unfair and is strictly prohibited.

Driven by these various motivations, we will soon organize an open machine learning competition hosted at *knowledgepit.ai* platform. Based on data from *Tactical Troops: Anthracite Shift* provided by us, the task will be to create a model that differentiates human players from bots. The method presented in this article will serve as a baseline.

REFERENCES

[1] Maria Arinbjarnar and Daniel Kudenko. 2012. Actor Bots. In *Believable Bots: Can Computers Play Like People?*, Philip Hingston (Ed.). Springer-Verlag, Berlin, 69–97. https://doi.org/10.1007/978-3-642-32323-2_3

[2] Christian Arzate Cruz and Jorge Adolfo Ramirez Uresti. 2018. HRLBÂ²: A Reinforcement Learning Based Framework for Believable Bots. *Applied Sciences* 8, 12 (2018), 2453.

[3] Christine Bailey, Jiaming You, Gavan Acton, Adam Rankin, and Michael Katchabaw. 2012. *Believability Through Psychosocial Behaviour: Creating Bots That Are More Engaging and Entertaining.* Springer Berlin Heidelberg, Berlin, Heidelberg, 29–68. https://doi.org/10.1007/978-3-642-32323-2_2

[4] Anton Bogdanovych, Tomas Trescak, and Simeon Simoff. 2015. Formalising Believability and Building Believable Virtual Agents. In *Artificial Life and Computational Intelligence: First Australasian Conference, ACALCI 2015, February 5-7, 2015. Proceedings 1.* Springer, Newcastle, NSW, Australia, 142–156.

[5] Anton Bogdanovych, Tomas Trescak, and Simeon Simoff. 2016. What Makes Virtual Agents Believable? *Connection Science* 28, 1 (2016), 83–108.

[6] Bobby D. Bryant and Risto Miikkulainen. 2006. Evolving Stochastic Controller Networks for Intelligent Game Agents. In *2006 IEEE International Conference on Evolutionary Computation.* 1007–1014. https://doi.org/10.1109/CEC.2006.1688419

[7] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep Blue. *Artificial Intelligence* 134, 1-2 (2002), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

[8] Marc Cavazza. 2000. Al in Computer Games: Survey and Perspectives. *Virtual Reality* 5 (2000), 223–235.

[9] Peter I Cowling, Edward J Powley, and Daniel Whitehouse. 2012. Information Set Monte Carlo Tree Search. *IEEE Transactions on Computational Intelligence and AI in Games* 4, 2 (2012), 120–143.

[10] Peter I Cowling, Colin D Ward, and Edward J Powley. 2012. Ensemble Determinization in Monte Carlo Tree Search for the Imperfect Information Card Game Magic: The Gathering. *IEEE Transactions on Computational Intelligence and AI in Games* 4, 4 (2012), 241–257.

[11] Jacques Ehrmann, Cathy Lewis, and Phil Lewis. 1968. Homo Ludens Revisited. *Yale French Studies* 41 (1968), 31–57.

[12] Salma Elsayed and David J King. 2017. Affect and Believability in Game Characters: A Review of the Use of Affective Computing in Games. In *GAME-ON'2017, 18th annual Conference on Simulation and AI in Computer Games.* EUROSIS, 90–97.

[13] David Gamez, Zafeirios Fountas, and Andreas K Fidjeland. 2012. A Neurally Controlled Computer Game Avatar with Humanlike Behavior. *IEEE Transactions on Computational Intelligence and AI in Games* 5, 1 (2012), 1–14.

[14] Sylvain Gelly, Levente Kocsis, Marc Schoenauer, Michèle Sebag, David Silver, Csaba Szepesvári, and Olivier Teytaud. 2012. The Grand Challenge of Computer Go: Monte Carlo Tree Search and Extensions. *Communications ACM* 55, 3 (March 2012), 106–113. https://doi.org/10.1145/2093548.2093574

[15] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1999. The Belief-Desire-Intention Model of Agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5.* Springer, 1–10.

[16] Simon Hecker, Dengxin Dai, Alexander Liniger, Martin Hahner, and Luc Van Gool. 2020. Learning Accurate and Human-Like Driving Using Semantic Maps and Attention. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE, 2346–2353.

[17] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. 2018. Rainbow: Combining Improvements in Deep Reinforcement Learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32. 3215–3222.

[18] Philip Hingston. 2009. A Turing Test for Computer Game Bots. *IEEE Transactions on Computational Intelligence and AI in Games* 1, 3 (2009), 169–186.

[19] Carroll E Izard. 2013. *Human Emotions.* Springer Science & Business Media.

[20] Magnus Johansson. 2013. *Do non player characters dream of electric sheep?: A thesis about players, npcs, immersion and believability.* Ph.D. Dissertation. Department of Computer and Systems Sciences, Stockholm University.

[21] Igor V Karpov, Jacob Schrum, and Risto Miikkulainen. 2012. Believable Bot Navigation via Playback of Human Traces. In *Believable Bots: Can Computers Play Like People?* Springer, 151–170. https://doi.org/10.1007/978-3-642-32323-2_6

[22] Ahmed Khalifa, Aaron Isaksen, Julian Togelius, and Andy Nealen. 2016. Modifying MCTS for Human-Like General Video Game Playing. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16).* AAAI Press, 2514–2520.

[23] Leslie Kurke. 1999. Ancient Greek Board Games and How to Play Them. *Classical philology* 94, 3 (1999), 247–267.

[24] John Laird. 2002. Research in Human-Level AI using Computer Games. *Commun. ACM* 45 (01 2002), 32–35. https://doi.org/10.1145/502269.502290

[25] John Laird and Michael VanLent. 2001. Human-Level AI's Killer Application: Interactive Computer Games. *AI Magazine* 22, 2 (Jun. 2001), 15–26. https://doi.org/10.1609/aimag.v22i2.1558

[26] Michael Sangyeob Lee and Carrie Heeter. 2012. What do you mean by believable characters?: The effect of character rating and hostility on the perception of character believability. *Journal of Gaming & Virtual Worlds* 4, 1 (2012), 81–97. https://doi.org/10.1386/jgvw.4.1.81_1

[27] Lars Lidén. 2003. Artificial Stupidity: The Art of Intentional Mistakes. *AI Game Programming Wisdom* 2, 5 (2003), 41–48.

[28] Daniel Livingstone. 2006. Turing's Test and Believable AI in Games. *Computers in Entertainment (CIE)* 4, 1 (2006), 6–19. https://doi.org/10.1145/1111293.1111303

[29] Jacek Mańdziuk and Przemysław Szałaj. 2012. *Creating a Personality System for RTS Bots.* Springer Berlin Heidelberg, Berlin, Heidelberg, 231–264. https://doi.org/10.1007/978-3-642-32323-2_10

[30] Nuno Marques, Francisco Melo, Samuel Mascarenhas, Joao Dias, Rui Prada, and Ana Paiva. 2013. Towards Agents with Human-Like Decisions Under Uncertainty. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35. 2978–2983.

[31] John McCarthy. 1990. Chess as the Drosophila of AI. In *Computers, Chess, and Cognition*, T. Anthony Marsland and Jonathan Schaeffer (Eds.). Springer, 227–237.

[32] Stephanie Milani, Arthur Juliani, Ida Momennejad, Raluca Georgescu, Jaroslaw Rzepecki, Alison Shaw, Gavin Costello, Fei Fang, Sam Devlin, and Katja Hofmann. 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.* 1–18.

[33] Ian Millington. 2019. *AI for Games.* CRC Press.

[34] Maximiliano Miranda, Antonio A. Sánchez-Ruiz-Granados, and Federico Peinado. 2017. Pac-Man or Pac-Bot? Exploring Subjective Perception of Players' Humanity in Ms. Pac-Man. In *Conference of the Spanish Association for Videogames Sciences.* 1–12. https://api.semanticscholar.org/CorpusID:45425351

[35] James H Moor. 1976. An Analysis of the Turing Test. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 30, 4 (1976), 249–257.

[36] Takahiro Morita and Hiroshi Hosobe. 2023. Video Game Agents with Human-like Behavior using the Deep Q-Network and Biological Constraints. In *Proceedings of ICAART 2023*, Ana Paula Rocha, Luc Steels, and Jaap van den Herik (Eds.), Vol. 3. 525–531. https://doi.org/10.5220/0011697500003393

[37] Negin Nejati, Pat Langley, and Tolga Konik. 2006. Learning Hierarchical Task Networks by Observation. In *Proceedings of the 23rd international conference on Machine learning*, William W. Cohen and Andrew W. Moore (Eds.). 665–672.

[38] OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. arXiv:1912.06680 https://arxiv.org/abs/1912.06680

[39] Steven Rabin. 2013. *Game AI Pro: Collected Wisdom of Game AI Professionals.* CRC Press. ISBN=978-1466565968.

[40] Adam Rankin, Gavan Acton, and Michael Katchabaw. 2010. A Scalable Approach to Believable Non Player Characters in Modern Video Games. In *Proceedings of 11-th International Conference on Intelligent Games and Simulation*, Alladin Ayesh (Ed.), Vol. 2010. 8.

[41] Arthur L Samuel. 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development* 3, 3 (1959), 210–229. https://doi.org/10.1147/rd.33.0210

[42] Jonathan Schaeffer, Neil Burch, Yngvi Bjornsson, Akihiro Kishimoto, Martin Muller, Robert Lake, Paul Lu, and Steve Sutphen. 2007. Checkers is Solved. *science* 317, 5844 (2007), 1518–1522. https://doi.org/10.1126/science.1144079

[43] Jacob Schrum, Igor V. Karpov, and Risto Miikkulainen. 2012. *Human-Like Combat Behaviour via Multiobjective Neuroevolution.* Springer Berlin Heidelberg, Berlin, Heidelberg, 119–150. https://doi.org/10.1007/978-3-642-32323-2_5

[44] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529, 7587 (2016), 484–489. https://doi.org/10.1038/nature16961

[45] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (2017), 354–359. https://doi.org/10.1038/nature24270

[46] Bhuman Soni and Philip Hingston. 2008. Bots Trained to Play Like a Human are More Fun. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).* IEEE, 363–369. https://doi.org/10.1109/IJCNN.2008.4633818

[47] Maciej Świechowski. 2020. Game AI Competitions: Motivation for the Imitation Game-Playing Competition. In *2020 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Maria Ganzha, Leszek Maciaszek, and Marcin Paprzycki (Eds.), Vol. 21. IEEE, 155–160.

[48] Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. 2023. Monte Carlo Tree Search: A Review of Recent Modifications and Applications. *Artificial Intelligence Review* 56 (2023), 2497–2562. https://doi.org/10.1007/s10462-022-10228-y

[49] Maciej Świechowski, Daniel Lewiński, and Rafał Tyl. 2021. Combining Utility AI and MCTS Towards Creating Intelligent Agents in Video Games, with the Use Case of Tactical Troops: Anthracite Shift. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI).* 1–8. https://doi.org/10.1109/SSCI50451.2021.9660170

[50] Fabien Tencé, Cédric Buche, Pierre De Loor, and Olivier Marc. 2010. The Challenge of Believability in Video Games: Definitions, Agents Models and Imitation

Learning. In *Proceedings of GAMEON-ASIA'2010*, Wenji Mao and Lode Vermeersch (Eds.). 38–45.

[51] Julian Togelius, Georgios N Yannakakis, Sergey Karakovskiy, and Noor Shaker. 2012. *Assessing Believability*. Springer Publishing Company, 215–230.

[52] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354. https://doi.org/10.1038/s41586-019-1724-z

[53] Di Wang, Budhitama Subagdja, Ah-Hwee Tan, and Gee-Wah Ng. 2009. Creating Human-like Autonomous Players in Real-time First Person Shooter Computer Games. In *Proceedings of Twenty-First IAAI Conference*, Karen Haigh and Nestor Rychtyckyj (Eds.). 173–178.

[54] Henrik Warpefelt, Magnus Johansson, and Harko Verhagen. 2013. Analyzing the Believability of Game Character Behavior Using the Game Agent Matrix. In *Proceedings of DiGRA 2013 Conference*.

[55] Kevin Warwick and Huma Shah. 2016. Can Machines Think? A Report on Turing Test Experiments at the Royal Society. *Journal of experimental & Theoretical artificial Intelligence* 28, 6 (2016), 989–1007.

[56] Astrid Weiss and Manfred Tscheligi. 2012. *Rethinking the Human–Agent Relationship: Which Social Cues Do Interactive Agents Really Need to Have?* Springer Berlin Heidelberg, Berlin, Heidelberg, 1–28. https://doi.org/10.1007/978-3-642-32323-2_1

[57] Minerva Wu, Je Seok Lee, and Constance Steinkuehler. 2021. Understanding Tilt in Esports: A study on Young League of Legends Players. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, Yoshifumi Kitamura (Ed.). ACM New York, 1–9. https://doi.org/10.1145/3411764.3445143

[58] Lotfi A Zadeh. 2008. Toward human level machine intelligence-is it achievable? the need for a paradigm shift. *IEEE Computational Intelligence Magazine* 3, 3 (2008), 11–22.

[59] Laura Zuchowska, Krzysztof Kutt, and Grzegorz J Nalepa. 2021. Bartle Taxonomy-based Game for Affective and Personality Computing Research. In *Twelfth International Workshop Modelling and Reasoning in Context @ IJCAI*, Jörg Cassens, Rebekah Wegener, and Anders Kofod-Petersen (Eds.). 51–55.