Distinguishing Human and AI Gameplay in Atari Games: An Exploratory Study of Human Perception and Behavioral Classification

Toni Bingenheimer[1], Martin Weier[1][0000-0001-9685-5238], and Biying Fu[1,*][1111-2222-3333-4444]

1 Hochschule RheinMain – University of Applied Sciences and Arts, 65195 Wiesbaden, Germany
∗ Corresponding author: biying.fu@hs-rm.de

Abstract. This paper examines the extent to which the gameplay behavior of humans and AI-driven agents can be distinguished, and whether both human observers and machine learning methods are able to reliably identify these differences. As an experimental basis, we trained several Atari 2600 games using a reinforcement learning approach and conducted a quantitative study in which human participants were asked to classify gameplay clips according to whether they were played by a human, an AI agent, or a combination of both in single-player and multiplayer settings. The selected games come from the Atari 2600 environment, which is widely used as a benchmark in reinforcement learning research. Our study investigates not only whether such distinctions are perceivable, but also which visual or behavioral cues may influence judgments. The results contribute to a better understanding of human perception of AI behavior in interactive digital environments and offer implications for future work at the intersection of HCI, AI, and game-based evaluation methods.

Keywords: Game Pattern Recognition · Automated Pattern Recognition · Human vs AI.

1 Introduction

Artificial Intelligence (AI) plays an increasingly important role in digital applications and computer games [1]. At the same time, video games—particularly Atari 2600 titles—have long served as test environments for AI research and established benchmarks for reinforcement learning (RL), enabling the study of agent learning and decision-making [12].

Despite achieving, in some cases, superhuman performance, AI agents often exhibit gameplay behavior that differs from human play [3]. Human players typically demonstrate variability, errors, and context-dependent decisions, whereas AI agents tend to use highly optimized strategies that may not resemble human

behavior [17]. Whether these differences are perceptible is relevant to both AI research and the analysis of interactive systems.

This work investigates differences between human and AI-controlled gameplay in selected Atari games. Gameplay sequences from human players and RL agents were collected and evaluated in a user study, in which participants classified clips according to their presumed origin. In addition, we examine whether a trained classifier can distinguish between human and AI gameplay more reliably than human observers.

The study therefore addresses both human perception of AI behavior and the automated detection of behavioral patterns. It can be understood as a game-specific behavioral Turing test [5], in which participants judge whether observed gameplay appears human or AI-controlled rather than assessing intelligence through language.

2 Related Works

2.1 Reinforcement Learning and Atari Games

Reinforcement learning trains agents to select actions through interaction with an environment while maximizing cumulative reward [19]. Atari 2600 games are widely used as benchmarks because they provide diverse tasks with visual input, discrete actions, and defined reward structures. The Arcade Learning Environment offers a standardized platform for evaluating agents across multiple games [13].

Deep Q-Learning demonstrated that agents can learn control policies directly from game input and achieve strong performance across several Atari titles [14]. Later methods, such as Agent57, surpassed human performance on the Atari 2600 benchmark suite [4]. However, high task performance does not necessarily indicate human-like gameplay. Agents may use highly consistent and optimized strategies that differ from human behavior. Atari games therefore provide a suitable basis for examining not only how well AI agents play, but also whether their behavior can be distinguished from that of human players.

2.2 Human-like AI in Games

Digital games are important environments for studying artificial intelligence because they combine perception, decision-making, planning, real-time interaction, and situated behavior within complex but constrained settings [7]. Consequently, games can be used to evaluate not only performance, but also whether agents behave in understandable, believable, or human-like ways.

Human-like behavior is closely related to believability. A believable agent need not play optimally; it should behave plausibly within the game context. Human players commonly display variability, hesitation, errors, and context-dependent decisions, whereas AI agents may follow consistent and optimized strategies. Performance metrics alone are therefore insufficient when the goal is to create natural or human-like gameplay [2].

Believability is multidimensional and may involve movement, timing, responsiveness, goal-directedness, errors, and reactions to changing situations. The Game Agent Matrix emphasizes that believable behavior depends on the relationship between the agent, the game context, and player expectations [20]. In Atari-style games, even limited action spaces can reveal differences through reaction timing, action consistency, movement trajectories, and repeated decision sequences. This study therefore examines whether human observers and automated behavioral features can distinguish AI-controlled from human gameplay.

2.3 Behavioral Turing Test for Game AI

The idea of assessing AI by its ability to appear human originates from Turing's imitation game. In game AI, this concept has been adapted to evaluate whether an agent behaves like a human player rather than whether it can communicate like one [11]. The present study applies this idea as a behavioral Turing test for Atari gameplay. Participants view gameplay sequences and determine whether they were produced by a human or an AI agent. This measures human-likeness from a perceptual perspective by testing whether observable behavior reveals the source of the gameplay.

Human judgments may nevertheless overlook subtle temporal and sequential patterns, such as reaction-time variability, action-switching frequency, or repeated action sequences. Therefore, this study combines a user study with a behavior-based machine learning classifier. While human participants evaluate gameplay visually and holistically, the classifier analyzes quantitative features extracted from gameplay sequences.

3 Selection of Atari Games

For this study, five games were selected: *Pong*, *Snake* [1], *Breakout*, *Kaboom!*, and *Space Invaders*. The selection was made with the objective of covering different requirements regarding perception, reaction speed, planning, and decision-making. At the same time, several of the chosen games are among the established benchmark environments in RL research [12,15].

The selected games differ considerably in their gameplay characteristics. *Pong* and *Kaboom!* primarily represent reaction-based tasks that require fast and precise control decisions. *Breakout* extends these requirements by incorporating strategic elements, as players must not only control the ball but also systematically clear blocks. In contrast, *Snake* requires longer-term planning of movement sequences. Finally, *Space Invaders* constitutes the most complex decision-making environment within the selection due to the need to simultaneously manage multiple objectives and actions.

To generate AI-controlled gameplay behavior, we trained the different games by using RL methods. Depending on the complexity of the respective game environment, both tabular Q-learning approaches and Deep Q-Learning methods

[1] `https://github.com/patrickloeber/snake-ai-pytorch`

were utilized [15,18]. However, the focus of this work is not on comparing different learning algorithms, but rather on analyzing and identifying the resulting gameplay behavior.

By combining games with diverse requirements, a comprehensive experimental basis is established that encompasses both simple reactive behavior patterns and more complex strategic decision-making processes. This enables a differentiated analysis of whether human and AI-controlled gameplay can be reliably distinguished by both human observers and automated methods.

4 Methodology

To investigate the perceptibility of human and AI-controlled gameplay behavior, an experimental online study was conducted. Data collection was carried out using a standardized questionnaire implemented with SoSci Survey [9]. Participants were presented with short video sequences from the selected Atari games, which had been generated either by human players or by trained RL agents. Their task was to classify the origin of the observed gameplay behavior.

The study design comprised three different scenarios.

- In the *first* scenario, participants were shown individual videos depicting either human or AI-controlled gameplay.
- In the *second* scenario, two videos from the same game were presented side by side, allowing participants to directly compare both gameplay styles.
- In the *third* scenario, *Pong* matches involving two simultaneously acting players were examined.

Participants were asked to determine whether the players were human, AI-controlled, or a combination of both. We further asked participants to rate their confidence in their decisions using a five-point Likert scale.

In addition to the classification task, participant's subjective confidence was recorded for each decision. A five-point Likert scale ranging from "strongly disagree" to "strongly agree" was used for this purpose [10]. Furthermore, demographic information as well as self-assessments regarding video game experience and familiarity with artificial intelligence were collected.

The video material consisted of recorded gameplay sequences from human players and trained AI agents. To avoid selection bias, gameplay clips were sampled according to predefined criteria. For each game and agent type, clips were selected to cover comparable duration and gameplay phase. The selection process was performed before participant responses were collected. The sequences were selected to include both potentially distinctive and difficult-to-distinguish behavioral patterns. All participants were presented with identical video material under standardized conditions.

The data were analyzed descriptively. Classification decisions were coded as either correct or incorrect and evaluated in terms of recognition accuracy. In addition, means and standard deviations were calculated for both recognition performance and subjective confidence. Pearson's correlation coefficient[6] was

used to examine the relationship between perceived confidence and actual performance. Furthermore, the results were analyzed separately with respect to demographic characteristics, video game experience, and task format.

5 Evaluation results from the Human Recognition Study

A total of 78 individuals participated in the study. After excluding incomplete datasets, the responses of 52 participants were included in the analysis. The adjusted mean completion time was 15.27 minutes (SD = 9.71). We further evaluated the results by leveraging the recognition accuracy and the pearson correlation coefficient (r-score) [16].

Across all task formats, the mean recognition accuracy was 50.28% (SD = 14.9%). The average subjective confidence rating was 3.60 (SD = 0.57) on a five-point Likert scale. No meaningful relationship was observed between subjective confidence and actual classification performance (r = -0.047), suggesting that participants were only limitedly able to assess their own performance accurately. An overview of the classification accuracy across the different task formats is shown in Fig. 1.

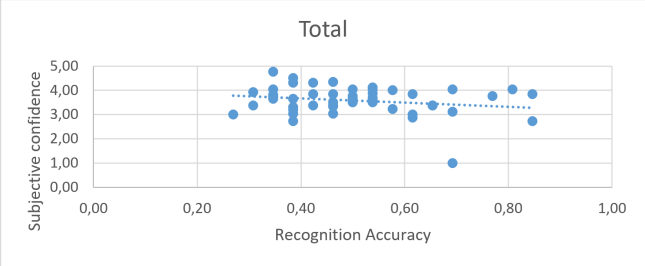
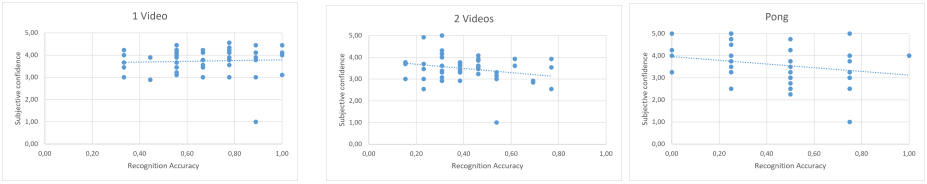


Fig. 1: depicts the recognition accuracy vs subjective confidence among all participants and scenarios.

The results differed substantially across the three task formats, as shown in Figures 2a, 2b, and 2c. The highest performance was achieved in the classification of individual videos, with a mean recognition accuracy of 65.58%. When two videos from the same game were presented simultaneously, recognition accuracy decreased to 40.75%. For *Pong* scenarios involving two simultaneously acting players, the mean recognition accuracy was 43.75%.

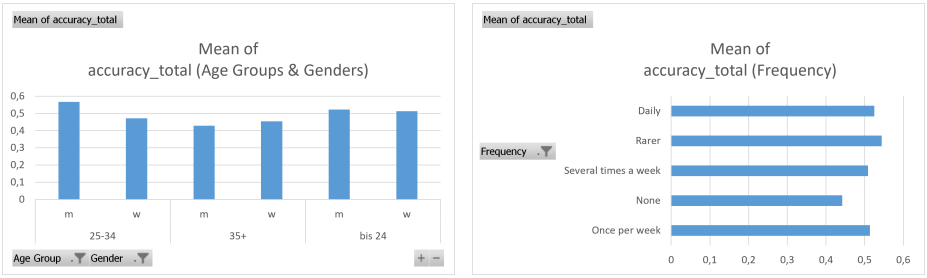
Interestingly, subjective confidence remained largely constant across all task formats, despite considerable differences in actual classification performance. While a weak positive relationship between confidence and performance was observed for individual videos (r = 0.125), the two more complex task formats showed negative correlations (r = -0.161 and r = -0.261, respectively). This suggests that higher subjective confidence was not necessarily associated with better classification performance.



(a) (AI or human), Scene 1 (b) 2 videos, Scene 2 (c) (Pong), Scene 3

Fig. 2: depicts the recognition accuracy vs subjective confidence in the task format for individual videos (AI or human) – Scenario 1 (Left), two parallel videos (left/right) – Scenario 2 (Middle) and *Pong* – Scenario 3 (Right).

Additional analyses by age, gender, and video game experience revealed only moderate differences between groups. The largest differences were found between participants with and without regular video game experience, whereas gender and age appeared to have only a limited influence on classification performance. Fig. 3a and Fig. 3b illustrate these results.



(a) (AI or human), Scene 1 (b) 2 videos, Scene 2

Fig. 3: Left shows the recognition accuracy and subjective confidence devided by age group and gender (m: male, w: woman) and Right based on video game experience.

We acknowledge that raw recognition accuracy is not strictly comparable across scenarios with different comparison criteria. Therefore, accuracy should be interpreted within the context of each specific condition rather than used for direct comparisons across scenarios.

6 AI-based Decision Making

In addition to the user study, an automated classifier was developed to investigate whether human and AI-controlled gameplay can be reliably distinguished

based on quantitative behavioral features. The objective was to compare the performance of a data-driven approach with the recognition performance of human observers.

For this purpose, gameplay sequences from human players and trained AI agents were recorded and segmented into fixed time windows. Various behavioral features were extracted from these sequences, describing temporal, action-related, and sequential characteristics of gameplay behavior. These features included, among others, the variability of reaction times, the frequency of action changes, and the length of consecutive action sequences. The selection of these features was based on the assumption that human and AI-controlled gameplay differ particularly with respect to consistency, predictability, and reaction dynamics.

A logistic regression model [8] was employed as the classification approach. The data were split strictly by players and gameplay sessions to prevent data leakage between the training and test sets. Predictions from multiple time windows were aggregated for each gameplay session, resulting in a final probability estimate of AI-controlled behavior for each gameplay sequence.

The same games used in the user study were also employed for model evaluation. Of the seven human gameplay sequences, six were correctly classified as human. Of the five AI-generated sequences, four were correctly identified. This corresponds to an overall classification accuracy of 83.3%. Fig. 4 shows the classifier's overall results for each game.

Compared to the user study, the classifier achieved substantially higher performance. While human participants reached an average recognition accuracy of 50.28% across all task formats, the model correctly classified the majority of gameplay sequences. These findings suggest that statistical behavioral patterns exist that can be detected more reliably by automated methods than by human observers.

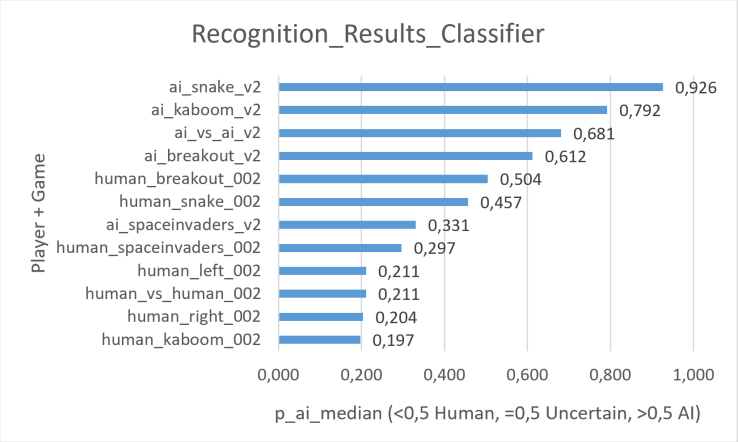


Fig. 4: shows the session-based classification results from the trained classifier.

7 Discussion

The results indicate that human observers had limited ability to distinguish between human and AI-controlled gameplay. Although performance exceeded chance level for some individual videos, accuracy declined in more complex tasks, while participants' confidence remained largely unchanged and showed no clear relationship with actual performance.

In contrast, the trained classifier achieved substantially higher accuracy, suggesting that human and AI gameplay differ in subtle behavioral patterns that are difficult to perceive consciously. Temporal decision consistency, reaction patterns, and action-sequence variability appeared particularly informative for automated detection. However, the limited number of sessions prevents conclusions about whether these patterns generalize across unseen players, games, and RL algorithms.

The findings further suggest that high-performing RL agents retain characteristic behaviors that differ from those of humans. Developing more human-like agents may therefore require not only optimizing performance but also modeling behavioral variability, errors, and context-dependent decision-making.

This study is limited by its small participant sample and limited gameplay data, as well as its focus on selected Atari games with simple, structured mechanics. Nevertheless, gameplay behavior analysis shows promise for game analytics, bot detection, and the development of more believable AI systems.

8 Conclusion

This work investigated how human and AI-controlled gameplay differs in selected Atari games and whether these differences can be identified by human observers and automated methods. In a subjective study, participants classified gameplay clips as being generated by human players, AI agents, or both in single-player and multiplayer settings.

Human observers achieved an average accuracy of 50.28 %, which was close to chance level, indicating limited ability to distinguish between human and AI-generated gameplay. The weak relationship between confidence and actual performance further suggests that participants tended to overestimate their classification accuracy. In contrast, the behavior-based classifier achieved 83.3 % accuracy in distinguishing human players from AI agents. These findings indicate that differences between the two groups are primarily reflected in temporal and action-related behavioral patterns, which automated methods can detect more reliably than visual observation alone.

Overall, the results show that RL agents exhibit characteristic behaviors despite their high performance. Behavior-based analysis therefore offers potential for player classification, bot detection, game analytics, and the development of more believable AI systems. Future work should evaluate larger datasets, additional game genres, and more advanced classification methods to assess the generalizability of these findings.

References

1. KI:Kultur – Wie künstliche Intelligenz in den Alltag findet (2025), https://ekw-verlag.de/ki-kultur/
2. Asensio, J.M.L., Peralta, J., Arrabales, R., Bedia, M.G., Cortez, P., Peña, A.L.: Artificial intelligence approaches for the generation and assessment of believable human-like behaviour in virtual characters. Expert Systems with Applications **41**(16), 7281–7290 (2014)
3. Badia, A.P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, D.Z., Blundell, C.: Agent57: Outperforming the atari human benchmark. In: Proceedings of the 37th International Conference on Machine Learning (ICML 2020). Proceedings of Machine Learning Research, vol. 119, pp. 507–517 (2020), https://proceedings.mlr.press/v119/badia20a.html
4. Badia, A.P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, Z.D., Blundell, C.: Agent57: Outperforming the atari human benchmark. In: International conference on machine learning. pp. 507–517. PMLR (2020)
5. Daniel, L.: Turing's test and believable ai in games. Computers in Entertainment **4**(1), 6 (01 2006). https://doi.org/10.1145/1111293.1111303, https://cir.nii.ac.jp/crid/1361418520462167424
6. Emerson, R.W.: Causation and pearson's correlation coefficient. Journal of visual impairment & blindness **109**(3), 242–244 (2015)
7. Laird, J., VanLent, M.: Human-level ai's killer application: Interactive computer games. AI magazine **22**(2), 15–15 (2001)
8. LaValley, M.P.: Logistic regression. Circulation **117**(18), 2395–2399 (2008)
9. Leiner, D.J.: Sosci survey (version 3.1. 06). Computer software]. https://www.soscisurvey. de (2019)
10. Likert, R.: A technique for the measurement of attitudes. Archives of Psychology **22**(140), 1–55 (1932), https://legacy.voteview.com/pdf/Likert_1932.pdf
11. Livingstone, D.: Turing's test and believable ai in games. Computers in Entertainment **4**(1), 6 (2006)
12. Machado, M.C., Bellemare, M.G., Talvitie, E., Veness, J., Hausknecht, M., Bowling, M.: Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. arXiv preprint arXiv:1709.06009 (2017), https://arxiv.org/pdf/1709.06009.pdf
13. Machado, M.C., Bellemare, M.G., Talvitie, E., Veness, J., Hausknecht, M., Bowling, M.: Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. Journal of Artificial Intelligence Research **61**, 523–562 (2018)
14. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)
15. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013), https://arxiv.org/pdf/1312.5602.pdf
16. Pearson, K.: I. mathematical contributions to the theory of evolution.—vii. on the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **195**(262-273), 1–47 (1900)
17. Seif El-Nasr, M., Drachen, A., Canossa, A. (eds.): Game Analytics: Maximizing the Value of Player Data. Springer London (2013). https://doi.org/10.1007/978-1-4471-4769-5, https://doi.org/10.1007/978-1-4471-4769-5

18. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, 2 edn. (2018), https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf
19. Sutton, R.S., Barto, A.G., et al.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)
20. Warpefelt, H., Johansson, M., Verhagen, H.: Analyzing the believability of game character behavior using the game agent matrix. In: Proceedings of DiGRA 2013 Conference (2013)