Comparing the Trade-off of Believability and Performance of Abstract Intelligent Agents and Humans Playing Super Mario Bros

Edward Morgan and Konstantinos Papangelis[(✉)]

Cardiff School of Computer Science and Informatics,
Cardiff University, Cardiff, UK
{MorganEH4,papangelisk}@cardiff.ac.uk

Abstract. In this paper, we will examine how the performance and player believability aspects of abstract intelligent agents within video games interact and affect one another. We will present the study used to assess whether performance and believability do have an effect on one another and how much of an effect. Followed by the results to the study and a discussion on potential design approaches to produce more believable agents that still provide the level of effectiveness expected of them.

Keywords: Believability · Intelligent agents · Video games

1 Introduction

With games getting increasingly complex in terms of their size and capabilities comes an expanded amount of focus on artificial intelligence and the usage of abstract intelligent agents within games. The increased computing power that comes with gaming platforms now means that agents can be used more actively to boost player enjoyment and immersion within their gaming experience [1, 2].

In more complex games such as Role Playing Games (RPGs), First Person Shooters (FPS) and Massive Multiplayer Online games (MMOs), where players interact extensively with agents there is an opportunity to further the player's enjoyment and immersion by providing them with agents that are both effective and that contribute an experience that feels like you are interacting with other humans [1, 2].

The aim of this study is to explore the aforementioned and provide initial evidence on how performance of an agent affects the player believability of said agent. Player believability in this context has been defined as "Someone believes that the player controlling the character/bot is real, i.e. that a human is playing as that character instead of the character being computer-controlled" [3].

2   Methods of Research

The research in this study was conducted using Super Mario Bros as a benchmark for both the performance and believability aspects. Four human players and four agents were recorded playing a series of identical levels. They played three different levels of four different difficulty settings (Twelve trials in total).

The four agents use different algorithms and approaches to play the game. We will be exploring agents that use an A* search algorithm approach, a rule based approach, and two simple agents - one that attempts to always move to the right of the screen jumping whenever it detects an enemy or an obstacle within a set distance of itself, while the other one attempts to constantly move to the right, jumping at every available chance. The A* search algorithm and Rule based approach are both common methods used in game agents and are well documented. The two simple agents are likely to perform worse than the two other agents but should offer insight into whether the performance gap affects their player believability. The human player video clips are used as a control variable.

The performances of each player (both human and agent) were tracked throughout the trials to give an overall performance score. This score can then be used to compare the players. The criteria for the performance were, Level Completion, Total Kills, Mario Status, Time Left, and Mario Mode. These were recorded for each difficulty level and then a running total for the complete run (All 12 trials).

The believability portion of the research was done in two parts by using videos of each player's run through and showing them to third party observers. The first part was to show the videos to single observers, asking them after each clip whether they believed it was a human or an agent, to put how sure they were on a scale (1 being 100 % sure it was an agent, 5 being unsure and 10 being 100 % sure it was a human.), and what skill level they believed the player was (Novice, Intermediate, or Advanced). During the clips the observers would talk out loud about what they were thinking allowing for us to take notes and identify certain traits that influenced their decisions on believability. This was done for 10 interviewees.

The second part of the research was similar to the first but differed in the fact that it was done with a large group (30 observers) at once, they filled out a questionnaire as the clips were shown. As it was such a large group it was impractical for us to probe for thoughts while the video clips were being shown. Using a large group to perform research at once advanced things considerably and allowed for more statistical data to analyse.

The research participants were a mixture of males and females and were all between the ages of 18-25. The group of participants had varying degrees of video game experience.

3   Analysis

The results gained from the research discussed above can be seen in Tables 1 and 2 below. Table 1 presents the results for the agent players, and Table 2 shows the results for the human players. These tables show the number of judges that guessed whether

Table 1. Agent player result table

	1. Forward Agent	2. Rule Based Agent	3. A* Agent	4. Forward Jumping Agent
5. **Performance Score**	8676.36	9865.94	11675.67	9187.99
6. **Human Votes**	3	15	0	14
7. **Agent Votes**	37	25	40	26
8. **Believability Score**	2.4	3.9	1.2	3.8

Table 2. Human player results table

	9. **Human 1**	10. **Human 2**	11. **Human 3**	12. **Human 4**
13. **Performance Score**	10339.33	9113.51	11587.67	9885.19
14. **Human Votes**	20	21	21	37
15. **Agent** 16. **Votes**	20	19	19	3
Believability Scale	6.6	5.3	4.9	8.1

each player was human or agent and the average of the believability scale used along with their performance score.

From Tables 1 and 2 we can see that overall, the performance did have an effect on the believability scores. However this was not the sole reason for the scores as discovered from the interviewees' comments. Examples of this were the forward agent who scored a low believability score as interviewees said it was so unintelligent it was obviously an agent regardless of the performance. Similarly observers were largely

unsure of human 2 as they played in a human fashion along with unintelligent behaviour displayed occasionally throughout.

The A* Agent which had the highest performance scored a 1.15 average believability which implies the observers were very sure that It was an agent. While the Human 3 had only a slightly worse performance score but the observers were unsure whether or not it was an agent or a human with a score of 4.9. Comments made by the interviewees said that the speed, and the general high performance level of Human 3 led them to believe it was an agent but then at higher difficulties it was offset by the fact that there was slight mistakes, taking some non-optimal paths, and occasionally had periods of waiting (For decision making or for enemies to clear).

The only human who posted a somewhat sure 8.1 believability score was the slowest and had the third lowest performance score. This suggests that the observers were looking for anything that made them seem agent-like such as being repetitive, methodical, and fast whether it had a good or bad effect along with optimality and drew conclusions. This was because they knew the goal was to identify agents and didn't know how many of the eight players were in fact agents. However all the agents still scored lower on the scale.

Unintelligent actions were related to jumping straight into danger or holes, ducking on screen when flying enemies were present although offered no obvious danger and not going backwards to avoid enemies.

4   Discussion and Conclusion

Bearing in mind that performance has an effect on the player believability of an agent; it brings us to how an agent should be designed to provide a player believable experience. The results imply that the correct level of performance has to be found to influence believability in the correct way. If the agent becomes too effective then they will come across as unhuman-like, although unintelligent decisions also suggest unhuman-likeness. This is shown from the interviewees comments made from the research.

Bungie, the developers of Halo talk about how tougher (More health points) agents leads to players believing they were smarter and more humanlike due to increased exposure and providing more of a challenge than agents who are actually more complex but provide less of a challenge due to being overcome quicker [5]. This of course was not their only method to provide versatile agents and has to be done in tangent with others.

While Togelius et al. [3] suggest that this should be done through a mixture of algorithm and level design optimization. This is an interesting suggestion that the level design could heavily influence the believability of agents [3]. Perhaps an agent could be designed to perform 'slower' at levels of lower difficulty opposed to performing at the highest speed capable giving a more human like feel. This change wouldn't necessarily make the agent perform worse, just at a more human manageable pace.

Using this in line with level design theory talked about by Jeremey Parish [4], could provide players with both a forgiving yet rewarding learning curve and well-rounded agents. It is conceivable that as a player progresses through a game they will become

more proficient and therefore become faster at actions. Consequently as a player becomes more proficient then the agents they encounter can too become faster.

Another suggestion is to use a heavy observation learning and case-based reasoning approach to produce believable behaviour [6]. This mixed with reinforced learning to produce a high level of performance could have the potential to produce these well-rounded agents we've discussed [6].

Overall, In this study we have looked at how two important aspects of abstract intelligent agents interact and influence each other. The initial data illustrates that performance does affect the believability of an agent, and how this could possibly be addressed via both algorithm optimization as well as level design optimization.

References

1. Charles, D.: Enhancing Gameplay: Challenges for Artificial Intelligence in Digital Games. University of Utrecht, The Netherlands (2003)
2. Woyach, S.: Immersion through video games, Illumin, vol. 5, no. 4 (2004)
3. Togelius, J., Yannakakis, G.N., Karakovski, S., Shaker, N.: Assessing believability. In: Hingston, P. (ed.) Believable Bots, pp. 215–230. Springer, Berlin Hiedelberg (2012)
4. Parish, J.: 1Up, 13 August 2012. http://www.1up.com/features/learning-level-design-mario. [Accessed 31 March 2015]
5. Butcher, C., Griesemer, J.: The illusion of intelligence: the integration of AI and level design in halo. In: Game Developers Conference, San Francisco (2002)
6. Iskander, U., Mozgovoy, M., Clint Rogers, P.: Believable and effective AI agents in virtual worlds: current state and future perspectives. Int. J. Gaming Comput. Mediat. Simul. (IJGCMS) **4**(2), 37–59 (2012)