Studying Believability Assessment in Racing Games

Cristiana Pacheco
Queen Mary University of London
London, United Kingdom
c.pacheco@qmul.ac.uk

Laurissa Tokarchuk
Queen Mary University of London
London, United Kingdom
laurissa.tokarchuk@qmul.ac.uk

Diego Pérez-Liébana
Queen Mary University of London
London, United Kingdom
diego.perez@qmul.ac.uk

ABSTRACT

Believability is a hard concept to define in video games. It depends on how and what one considers to be "believable", which is often very subjective. In previous years, several researchers have tried to find ways of assessing such concepts in games through Turing Tests on agents, which were programmed to behave like a human instead of focusing only on winning. Examples are the Mario AI Competition and the 2K BotPrize. Given the small pool of explored parameters and a focus on programming the bots rather than the assessment, in this paper we present work examining different methods of evaluating believability in video games. We explore believability through recorded gameplay and allow judges to analyze it. However, we use different parameters - such as ranking rather than binary answers - for asking how human-like the presented behaviours are. The objective of this study is to analyze the different ways believability can be assessed, for humans and non-player characters (NPCs) by comparing how results between them and scores are affected in both when changing the parameters. In order to provide a more general analysis, the study is carried out using two different racing games rather than one. Results show that these parameters have indeed changed the overall results of the study and how important it is to be able to generalize these concepts in game AI, given how clear it is that believability is dependent on genre, game and even the design of the questionnaire.

CCS CONCEPTS

• **Software and its engineering** → **Interactive games**; • **Computing methodologies** → *Artificial intelligence*;

KEYWORDS

Believability Assessment; Turing Test; Believability; Human-like Playing, Racing Games

ACM Reference Format:
Cristiana Pacheco, Laurissa Tokarchuk, and Diego Pérez-Liébana. 2018. Studying Believability Assessment in Racing Games. In *Foundations of Digital Games 2018 (FDG18), August 7–10, 2018, Malmö, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3235765.3235797

FDG'18, August 7-10, 2018, Malmö, Sweden

ACM ISBN 978-1-4503-6571-0/18/08…$15.00
https://doi.org/10.1145/3235765.3235797

1 INTRODUCTION

Artificial Intelligence (AI) has been widely used in video games to create agents for non-player characters. This has compelled many researchers to study and develop algorithms and techniques to create agents to compete against/besides humans and aid/defeat them. Another field of study, which has received comparatively less focus, is the production of agents that exhibit human-like behaviour. Some research has gone into the development of tracks in game competitions by adapting the original Turing Test [25] and trying to evaluate how believable NPCs are [5, 24]. To compete in these tracks, programmers would try to generate bots that opt for acting like human beings, instead of focusing on winning.

Games seem to be more engaging to players when they believe they are facing/cooperating with other human players [27]; perhaps because people are less predictable and also capable of provoking emotion in others. This seems to show that having more human-like NPCs could easily bring advantages for certain games and, perhaps, create an interest for both researchers and developers to try and create them. But how to know if we are interacting with such agents?

Assessing the believability of bots is an important step and currently there is no standard format for such testing. Here, we focus on this concept; what it is, which parameters are available, and why we should use them. Previous studies present a diverse range of competitions, assessment parameters and evaluation, making it difficult to accurately distinguish which one had the most human-like bots.

The main contribution of this paper is to first show the differences in how several believability assessments have been carried out in the past, to then show (both qualitatively and quantitatively) how multiple factors such as player experience, judge experience and even the presentation of the study can affect the accuracy of the assessment. The work performed in this paper is, therefore, less concerned with evaluating the human-like playing abilities of the bots and humans and, instead focusing more on the methods required to arrive at a successful judgement. In this study, we have alternated between different parameters proposed before [7, 24] and evaluated an adapted Turing Test with several versions. Our findings show differences between scores and opinions. We have also tried to collect data in a way that allows for a more general assessment rather than a game-genre-focused solution. This should, hopefully, motivate future studies to concentrate on the design of the study before the actual assessment is carried out, not just for racing but for other genres of games also.

This paper is structured as follows. Section 2 explores the previous work in this area, to then describe our own study in Section 3. Section 4 describes the qualitative and quantitative results of this study, concluding with final remarks and ideas for future work in Section 5.

2 LITERATURE REVIEW

2.1 Believability

In order to create human-like characters in video games one should explore the concept of believability first. This is a broad notion by itself and, to the knowledge of the authors, still lacking a precise and generally accepted definition. Several possibilities are lightly explored by Tencé et al. [22] that range from broad definitions such as Thomas and Johnston - to whom character believability is about providing illusion of life [23] - to more fine-grained definitions such as Reidl's - where believability takes into account several elements such as emotions, personalities and intentions in order to give a character an "illusion of life" [15].

When considering these concepts specifically in the context of games, Livingstone [7] and Togelius et al. [24] have divided this concept into two classes: *Character Believability* and *Player Believability*. The first considers a fully autonomous agent/bot which gives the illusion of life. That is, there is no human controlling them, but the agent acts in a believable way to a human observer. Player believability means that a character gives the illusion that a human player is controlling the agent, rather than the computer. This is the definition that has been considered for this paper.

2.2 Assessment Parameters

Given the complexity of the term "believable" in the context of character creation, evaluation becomes an important and complicated step. Unfortunately, there is still no accepted method of evaluating believability. However, previous researchers have tried to address this essential step. Togelius et al. [24] have proposed several parameters that should be taken into account for the development of an assessment. For this paper, we are considering only the subjective assessment, which is when the observers' experience is reported through questions. Other methods such as objective (collection of data on players' physiological responses) [10, 28] and gameplay-based reports (tracking of gameplay statistics from players' experiences) [13] are beyond the scope of this paper. We shall now proceed to analyze each possibility within *subjective* assessment.

Type of Response. For this option, only two possibilities are presented. Allowing players a *free response* or choosing among predefined answers (*forced data*). The first may contain richer information but at the cost of increased difficulty in analyzing responses and questionable validity upon clustering speech/text. The latter, however, will provide data that is simpler to analyse and could provide accurate models for believable bots, albeit restraining to very specific information.

When. Should the questions be asked during gameplay? Or after the gameplay session? Which time frames should be considered?

How. When prompted with questionnaires, options are wide. However, three main question types can be identified: boolean (simple 'yes' or 'no' questions); ranking (offer a scale of some form) and preference (by offering comparisons of game sessions).

Judges' roles. For this section we are considering what is the observers' role during a study. For such, only two possibilities are presented; either a *first person* role or a *third person* role. While assessing in first person, the participant will be part of the gameplay itself. This means he or she will be playing while assessing whether the agents they are encountering are other players or not. When it comes to the third person assessment, the participant is simply observing the gameplay and not participating in it. He or she will then judge based on their observations.

Length of Assessment. This is perhaps one of the least explored methods. To the knowledge of the authors, there is no defined time interval that is optimal to judge believability. It is, however, another parameter that has to be considered - for how long should the sessions last?

Representation. The final parameter in our list takes into account how we represent believability. Should we treat it as a discrete variable (a set of defined states such as 'true' or 'false')? Or as a continuous one (number between two values such as percentages)? Or even decide if one continuous variable should be used or more than one? After all, emotional dimensions [3, 16] have been used in previous studies to represent different emotional states and how these affect players [17].

Prior to this, other parameters have been pointed out in [4, 7, 9]. These provide yet more guidelines for assessment such as combining human and non-player characters - to judge both players and bots believability for referencing, given that humans will also not provide a perfect score, the spectators' cultural origins which may also affect their decisions and, finally, the experience level of these - more experienced players would understand better the game and distinguish patterns with more ease, as mentioned in [2].

2.3 Turing Test

Perhaps the most familiar and recognized test for evaluating believability in a machine is the Turing Test [25]. The original test proposes making a person hold a text based conversation with both another human and a machine and, after a given period of time, check if this participant can distinguish which one is a bot. If the person cannot tell the difference between both, then the machine is deemed intelligent.

The idea that this would recognize machine intelligence has been refuted before - most noticeably by Searle [18]; and it would not be fully applicable in a video game setting. For such, we would have to make the concept broader and adapt to our needs, especially taking into consideration the previous parameters. Luckily, attempts into assessing believability by creating a Turing Test for games have been explored already.

2.3.1 2k BotPrize. The first attempt at assessing believability discussed in this paper shall be the 2K BotPrize Competition. This was an open competition that started in 2008 for the commercial game Unreal Tournament 2004 (UT2004) [5], where the objective was for programmers to create bots to behaved in the most human-like way possible. They would have to be able to fool the judges into thinking that the agent they were observing was being controlled by an actual player.

UT2004 is a first-person shooter (FPS) developed by Epic Games and Digital Extremes which allows for multi-player matches, where bots could easily be introduced given the simple interface. A server will simulate a virtual world (map) in which people can play in and/or programmed bots can run. The objective is for several players/bots

to kill other players by shooting them. However, for the competition, one of these matches was created with three participants only - a judge, a player and a bot. They were asked to play for a few rounds that lasted approximately 10 minutes. Players were to play as they would with any other game and for a small prize; judges had to rate their competitors using a given scale from 1 to 5 (where 1 meant that the player was 'not very human-like bot' and 5 meant that the player 'is human'). Names - for both humans and bots - were generated automatically after entering the game to 'player' and a number; physical appearance was kept as the same for all.

In this competition there were five judges, another five players and also five bots. To win, a bot would have to fool 4 of these judges into thinking it was a human. The final results show that only two bots managed to deceive two judges and another deceived only one. Players always came out as more human than the bots and three of them passed the test. This means that judges managed to distinguish between player and bot with some players winning but no bots winning.

It ran once more in 2009 but with a particular change: the new test included 'special weapons' which were put to challenge the bots - they would provide an effect whose description was not given to the competitors such as 'freezing' or 'teleporting'. Judging, however, remained the same and the results were similar to 2008.

In later competitions, Hingston proposed making judging "part of the game"[6]. He provided players with yet another 'special weapon' to use against other players/bots. Given that in UT2004 guns have two different 'modes' - primary and alternate - the primary was supposed to be used against bots and the secondary was supposed to be used against players. If the player made the right decision (hit a bot with the primary mode or a player with the alternate) it would receive 10 points - which is much higher than the original 1 point for a simple frag. However, if the player made the wrong choice, it would instantly die and lose 10 points. Furthermore, names were then simply taken from an existing database with common names (ease in remembering) and the mod's details were fully available.

As for results, how 'human' characters were was defined by the number of human judgements divided by the total judgements, then multiplied by 100. Unfortunately, all the entries showed that humans came in all the top positions with the most human-like bot still not being able to beat the least human-like player.

2.3.2 Mario AI Championship. Another large competition that is relevant for believability assessment is the Mario AI Championship. In this case, the game chosen for the challenge was a modified version of the game Infinite Mario Bros - which is a clone of the game Super Mario Bros that belongs to Nintendo and is currently publicly available. It was modified for research purposes and details of these changes are available in [20]. Infinite Mario Bros is a 'platformer' in which the gameplay consists of a player controlling an agent (named Mario) on a 2D level. It is possible to make this character jump, move left and right, run, duck and fire (when in the appropriate state). The player can also interact with objects in the game (magic mushrooms, coins, green pipes, ...) and should avoid enemies or defeat them by jumping on them.

This championship started running from 2009 and it includes several tracks. However, in 2010, Togelius et al.[24] added a Turing Test Track. The idea behind this track was to ask contestants to program agents for the game that can play like a human being (giving the notion that 'Mario' is being controlled by a player) instead of playing as well as possible - which was the point of the other existing tracks.

To evaluate these agents, their gameplay was recorded before the public event and so was one of a human player. These videos were then randomly selected and presented to an audience who were asked to vote if they thought they were watching a player controlling the character or AI. If in doubt, they could simply answer "Not decided". The player was voted the most human-like with a bot coming second with 5 votes less.

In 2012, Shaker et al. [19] decided to make some changes to the evaluation method. Videos were once more recorded of both the agents and of two human players (inexperienced and expert) during their gameplay sessions. The AI bots and the players played three different levels of various difficulty - 0, 1 and 2 - where 0 would only offer mushroom-like enemies (the easiest kind to defeat); 1 introduced the turtle-like enemies (slightly harder to defeat/avoid) and gaps on the platforms; and 2 had an increase of the amount of enemies, which also included the flying kind (ever harder to defeat), and wider gaps.

This time, the videos were uploaded to a web server whose interface also allowed not only the viewing but also the collection of evaluations from the participants - after being given instructions on how to proceed with the questionnaire. Participants were asked to compare two randomly selected one-minute videos of different agents sessions. For this judgement, viewers were asked which one played in a more human-like manner and which one did they think was more expert. For the answers, four possibilities were given: one video or the other, both equally or none of them. Once again, a player won (novice - score of 30.95%) with a bot coming second (25.79%), beating, however, the expert player (23.21%).

2.3.3 Simulated Car Racing Championship. This championship is yet another competition to which agents have to be submitted that drive a car around given tracks. For this championship the game chosen was The Open Racing Car Simulator (TORCS) - a 3D car racing simulator created by Eric Espié and Christophe Guionneau with a realistic and complex physics engine that provides several cars and tracks for the players. In this game, several aspects are taken into account that would not usually be present in commercial video games, such as grip of the wheels, fuel consumption and even aerodynamics of the car itself.

For this championship, there were three 'phases' and in each phase the submitted controller had to race in three different tracks, unknown to the participant. In these tracks, which could be either new ones or already existing in game, there were also three different stages: a warm-up - to allow agents to redefine their car's parameters and adapt to the track in question by being allowed to race for 100000 game ticks, that would be equivalent to around 30 minutes of game time - qualifying - where the 8 best agents are chosen to race after being given the opportunity of performing 10000 game ticks alone in the track which would be the equivalent of around 3 minutes and 20 seconds - and, finally, the race - where these 8 compete with each other and given a score just like in Formula 1. A more detailed description can be found in [8]

Figure 1: 'Helicopter' perspective from game replay.

Figure 2: *Speed Dreams* Screenshot

Even though this championship was not developed for the single purpose of creating believable agents (or held any component of it), there were worthy attempts at creating a controller for this competition that would mimic a human driving style. Muñoz et al. [11] focused merely on an agent that could drive as a human and not necessarily win the race or pass the qualification phase. However, it produced noteworthy results and even managed to win in some races at a later attempt [12].

3 STUDY DESIGN

This section presents the motivations and design behind our study. We shall start with why it is important to develop such a study, show which previous work has been accomplished so far in assessing bots in different genres, which games are being used by this particular study and, finally, our own procedure for assessment.

One of the championships praised for its ingenuity was the 2k BotPrize [24]. Even though it got significant attention, we believe that making the assessment part of mechanics of the game was not optimal. We wanted to avoid a shift of priorities regarding the participants' tasks for this study and, given that the games being used are racing games, it would be harder to introduce mechanics that would allow players to 'flag' others if they believed them to be bots. 'Looking' at other cars, nudging or any other sort of force applied to them would affect the players' driving abilities and, unfortunately, make it obvious to other players. Not providing rewards does not give an incentive for participants to try to judge others, however, providing points for judging correctly or not could also give more reasons for players to behave 'poorly' on purpose. To avoid such behaviours, one could ask at the end of each race who was a bot or not in their opinion but this is still a challenge to one's memory even for short races. This made us come to the decision of first recording videos of gameplays and second, showcase them to a new set of participants. This would allow the players to focus fully on playing the game and the judges to focus fully on judging the game sessions. Nevertheless, we wished to introduce yet another variable in the videos perspective - the camera type - since *Forza Motorsport* allows for a 'helicopter' view (different type of third person game perspective displayed in Figure 1) in its *spectator* mode from saved replays. This would possibly show how a camera less common to the perspectives used by players could affect results positively or negatively.

We also chose to check the affect of the length of assessment and the prior experience of each participant and whether they influenced the scores or not. In the Mario AI Championship, videos were showcased of an agent's game session through a level. However, different time intervals were not tested and neither was the experience of the judges asked. In our study, participants were asked about their experience level in racing games at the beginning of the questionnaire. To address the former point, two video lengths were employed in this study: a 12 second version and a 35 second version (almost three times the length). All of our 12-seconds videos are sub-sequences of the 35-second versions. In other words, the 35-second videos are a longer snippet of the 12-second ones. The objective is to analyse if the same video but with added time would alter the final evaluation of the judges. Introducing a video from a completely different section of the game session would not be a fair test. For example, 12 seconds of the last lap where the player is about to finish first without any opponents intruding *VS* 35 seconds of the first lap with interactions between opponents would not be a fair comparison.

Finally, a third dimension is proposed in this study: the usage of two different questionnaire layouts. One of them with one video per page with the related questions, and the other with two side-by-side videos (similar to the Mario AI Championship), with the respective questions underneath each. The objective of this dimension is to investigate if showing one video or two side-by-side videos affects the precision of the judges, as it is possible that participants compare the videos when making their decisions.

3.1 The Games

Two games within the same genre (racing games) were used for this study. In this genre of games, the objective is to win races by controlling the car around a track using the keyboard or a console controller/joystick and finishing in the first position.

The reason for this choice is twofold: first, by focusing on two games rather than on one, we hope to draw conclusions that do not totally depend on the game used, following the spirit of the latest General Game AI trends [14]. Secondly, both games represent different populations in terms of audience (research versus commercial),

adding another dimension to the study presented in this paper. Shooters (*UT2004*) and platformers (*Mario*) have been used before for believability, but, to the knowledge of the authors, racing games (such as *TORCS*) have not. Two games were chosen for this study: *Speed Dreams* and *Forza Motorsport 6: Apex*, which are described next. While the former is mostly used in academic environments (*Speed Dreams* is an evolution of *TORCS*), the latter one has sold many copies commercially for both consoles and PC.

Speed Dreams[1] is a free and open source 3D car racing game written in C++ and developed by Gaëtan André. Its development started in 2008 and continued with several releases up until April 2016. It is a version of the TORCS simulator with more features, which include several racing modes and events that try to reproduce real life races such as championships and endurance. Being open source also means that it allows users and developers to create their own tracks and controllers which can be run within "Practice" mode. To gain points and contribute to the championship rankings a "Career" mode is also present for players to compete in several sessions with the provided car classes, random tracks and opponents.

It also features several realistic physics engines which take into account position, speed, damage and suspension with real-time car collisions. All of this will affect how the cars perform during races. It will also allows users to choose between different AI opponents, called "robots", that compute racing variables on their own - throttle, brake, steer, gearbox and clutch. For the players, however, there is the possibility of using driving aids. These include ABS, TCS and speed limiter for pit-stops. An sample screenshot can be seen in Figure 2.

Forza Motorsport 6: Apex[2] is another racing video game developed by Turn 10 Studios and published by Microsoft. The Apex version is the free-to-play version used for this study and it was published for the Windows 10 platform. This version was released in 2016, featuring many cars that can be driven on different tracks in several racing disciplines and courses. Forza has been placed within the simulation genre as well given that its cars' handling is programmed to behave as close to real-life as possible. These include tracks that were created out of real locations around the world with a few hand-made. Players are also allowed and encouraged to participate in championships where they will have to compete in several races based on vehicles' categories (power, drivetrain, vehicle age, class, etc.).

Perhaps the most important factor, which also compelled this game to be used in this study, is its bot behaviour. In this game, the NPCs are called "Drivatars" – trained AI based on players' data, recorded while racing in the game tracks and displayed in Figure 3. Even though learning by imitation has its flaws, for this game the benefits of mimicking human playing styles was higher and so, it was kept as a mechanic both in earlier and later versions of the game - [1, 26] describes how these work.

3.2 Procedure

To run this study, we have divided the procedure into two different parts: the recording and processing of videos, and the questionnaires for later judgment.

[1] http://www.speed-dreams.org/

[2] https://www.microsoft.com/en-gb/store/p/forza-motorsport-6-apex/9nblggh3shm7

Figure 3: *Drivatar* from *Forza Motorsport 6: Apex*

3.2.1 Videos. Several participants were asked to play both games, *Speed Dreams* and *Forza Motorsport 6: Apex*. Players were allowed to try the games before-hand until they felt confident to do the final and recorded race. The track they were given was always the same between all participants, with same difficulty level and conditions. The cars and respective colours were randomized between the same allowed class. In the end, players were asked to fill in a quick survey to gather information such as their experience level with racing games, how much did they enjoy each game and which controller (gamepad or keyboard) they used during recording[3]. For each game three videos of humans and three videos of bots (with different levels of experience) were recorded. For Speed Dreams this meant 6 videos in a third person perspective (TPP), however, for Forza, this meant 12 videos - 6 in TPP and another 6 in 'helicopter' view. The latter were recorded on the same time frame as the previous (same race point). Finally, per video, we took a 12 seconds cut and a 35 seconds cut (that includes the previous 12 seconds). These extracted videos were all around the same period, shortly after the start (to include interaction with adversaries). Having 6 videos for *Speed Dreams* used for both time versions and 12 videos of *Forza* also for both time versions equals to 36 videos in total. The videos from *Speed Dreams* were also edited to have names, driving aids and lap places/times covered with a black rectangle over them, so judges could only focus on driving abilities instead of other complementary factors.

Players were not told whether their videos would be used or not and their email addresses were also collected - we recorded several participants and only a few were selected. At this stage the authors had not made a decision on which to use - or even how they would be edited or displayed in the questionnaire. This made sure that the participants had very little information about the second part of the study and so, they would not be able to share information that could compromise the study. Furthermore, if the players accessed the link to do the second part of the study, emails could be compared to make sure they had not seen themselves play, in case their videos had been selected - given that the videos were randomly selected out of a large pool, it was also very unlikely that they would be presented with their own game sessions.

[3] Participants were asked to play with the controller they felt more comfortable with.

3.2.2 Questionnaires. The assessment type chosen was *subjective* (See Section 2.2), meaning that questions are presented to participants to collect their responses based on their appreciation of the videos they watch. The study then follows a very close procedure as Togelius et al. [24] - judging the believability of game sessions of previous participants and bots.

Having decided the type of questionnaire required, we move on to defining the required response format. For this parameter, we wanted to allow both types: simple multiple-choice to elicit measurable responses about particular aspects and free-form text response to collect richer data regarding the participants reasoning of why they answered in a particular way. Questionnaires were presented immediately after the participants watched the videos to minimise task intrusiveness and maximise their recall. The interface was kept as minimal as possible, consisting of multiple choice, direct questions, and slider bars. When it comes to the believability value, we decided to rank (i.e. treat the variable as a continuous value rather than a discrete one). The reason behind this is that a boolean, albeit simple, lacks data for further consideration; ranking, on the other hand, allows for a more fine grained analysis on the certainty of the judgements for both humans and bots.

The questionnaire was hosted online and 8 different questionnaires were possible, as combination of the following variables chosen at random for each new participant:

- **Type:** One video per page (type *Single*) or two videos per page, side by side (type *Pair*).
- **Length:** Short (12-seconds)[4] or long (35-seconds)[5] videos.
- **Camera:** Third-person perspective[6] or 'helicopter' camera[4]. As the Helicopter camera is only available in *Forza*, all *Speed Dreams* videos have third-person perspective.

Then, each questionnaire displays four videos from the six existent for each game, giving each participant a total of 8 videos to watch. All questionnaires started by asking participants about their experience in racing games, and then proceed to evaluate each one of the videos. For each video, the judges were asked to:

(1) Provide from 0 to 10, where 0 means they're confident they're dealing with bots, 5 it could be either and 10, definitely human.
(2) Optionally, provide a justification of their choice as free-form text.

Finally, all questionnaires finished with simple questions on preference over games to have more believable bots and ideas for future studies (for curiosity and improvement of the judgement experience). Participants knew only that this study would be about judging racing games, they were not told that the questionnaire had multiple variants; how many videos of each game were available; quantity of bots/humans in these videos or if they were right or wrong in their decision.

Overall, 110 participants with differing levels of experience in racing games took part in this study. Each one of them completed one of the 8 different types of questionnaires available, which was selected uniformly at random by the survey server. Table 1 summarizes the questionnaire types and how many times each one was presented to the judges.

[4]https://www.youtube.com/watch?v=GKTe7_EEHlM&feature=youtu.be
[5]https://www.youtube.com/watch?v=PVmCbmb_7-o&feature=youtu.be
[6]https://www.youtube.com/watch?v=aO1nT2Z4R7Y&feature=youtu.be

Table 1: Available questionnaire types and number of times each type has been presented to a judge. Each questionnaire asked for 8 videos to be evaluated, 4 for each one of the game, all selected uniformly at random from the video pool.

ID	Time (s)	Question	Camera	Sample Size
1	12	Single	Third	12
2	35	Single	Third	22
3	12	Pair	Third	10
4	35	Pair	Third	16
5	12	Single	TV	10
6	35	Single	TV	15
7	12	Pair	TV	11
8	35	Pair	TV	14

Table 2: Accuracy of the judge evaluations to identify bots and humans, with standard error. An answer is correct if the assessment is in the range $0-4$ for bots and $6-10$ for humans.

Parameters	Type	Game	
		Forza	**SPD**
Question	**Single**	57.63% (3.22)	58.47% (3.21)
	Pair	51.96% (3.50)	58.33% (3.45)
Camera	**TV**	53% (3.53)	55.5% (3.51)
	Third	56.67% (3.20)	60.83% (3.15)
Time	**12 sec**	51.74% (3.81)	55.23% (3.79)
	35 sec	57.01% (3.02)	60.45% (2.99)

4 RESULTS AND DISCUSSION

This section analyses the results of the questionnaires in two parts. First, the judging of videos: when participants were shown videos and were asked to both rate them and justify their answers. Second, the final two questions: the multiple choice question and the open ended question intended for any additional information. In the first part, the Chi-Square Statistical test was calculated aiming at a significance level of 0.05 to reject the null hypothesis.

4.1 Videos and Ranking

To represent the results, the data has been divided into different groups in order to compare the relevant variables and parameters previously discussed in Section 3. Tables 2, 3 and 4 show the percentage of correct answers by judges in evaluating if the driver in a video being watched is a bot or a human. The standard error of each value is given in brackets. The parameters in these tables are displayed

Table 3: Accuracy of the evaluations to identify bots and humans, with standard error, indicating the experience on racing games of the judges (No Experience, Some Experience and Experienced). An answer is correct if the assessment is in the range $0-4$ for bots and $6-10$ for humans.

Parameters	Type	Experience Level of the Judge (Sample size)			
		No Experience (28)	Some Experience (62)	Experienced (20)	All (110)
Question	Single	50% (5.33)	60.66% (2.96)	58.04% (4.66)	58.05% (2.27)
	Pair	48.53% (4.29)	56.25% (3.31)	68.75% (6.69)	55.15% (2.46)
Camera	TV	42.15% (4.37)	58.52% (3.71)	62.5% (4.94)	54.25% (2.49)
	Third	58.33% (5.03)	58.75% (2.75)	59.38% (6.14)	58.75% (2.25)
Time	12 sec	44.32% (5.30)	57.29% (3.57)	54.69% (6.22)	53.49% (2.69)
	35 sec	52.21% (4.28)	59.54% (2.82)	65.63% (4.85)	58.77% (2.13)

as 'Question' ('Single' or 'Pair'), 'Camera' ('TV'/'Helicopter' or 'Third') and 'Time' (12 or 35 seconds). Columns hold the names of the different groups, experience and games.

In order to decide what 'correct' means in this context, two different thresholds have been given for "correctness". The first, relaxed, threshold determines that the participant was correct in their judgement if they provided an answer in the range $0-4$ (inclusive) for a 'Bot' video or in the range $6-10$ (also inclusive) for a 'Human' video. The second threshold requires a higher confidence, with answers in the ranges $0-2$ and $8-10$ for correctness of judgement.

4.1.1 Games. Our first comparison addresses the differences in results based on the games. Even though the differences are not statistically significant, there seems to be an overall higher accuracy in SPD for all treatments considered, as seen in Table 2. A possible explanation for this is that Speed Dreams is a simulator with more complicated and realistic physics, according to the developers. This implies a higher difficulty in playing these games (more sensitive controls-wise) which in turn would produce more mistakes that can be identified by the judges. Bots are also built to race more in an optimal way rather than a way to please the player or even in a human-like manner. However, in the case of *Forza*, *Drivatars* are programmed based on players' actions and so are harder to judge.

The parameters have produced some interesting results: people that were presented with longer videos, in both games, had a greater number of correct answers than the ones presented with shorter videos. This specific trait was indeed expected as it is reasonable to think that longer videos allow for a more careful deliberation. However, results show that correctness is achieved in not much more than 50% of the cases, suggesting that even when given more than double of the time, it still might not have been enough to judge properly. In fact, it is interesting to note (as described below in more detail) that some judges have pointed out that 35 seconds was not enough time to judge. This also suggests that the optimal value for a gameplay session might prove difficult to find, taking into account each player and judge' capabilities.

Results show that the third perspective camera also yielded more accurate responses in both cases. Since players tend to play racing games with this perspective, it is possible to assume that it becomes easier to analyse the video with a more game-like view that is more familiar to them.

As for the Question variable, the 'Single' type produced higher accuracy in *Forza* and a similar performance in *Speed Dreams*. Another point many participants have made was the fact that they would compare the videos when they were in the same page without any instructions to do so; but other would ignore and do one at the time with no comparison. Therefore, it is possible to assume that this comparisons have led participants to incorrect decisions. It is clear that, at least in *Forza*, presenting isolated videos produces more accurate judgements. Evaluations are different using the same pool of videos, which were always randomly selected. This result is one of the examples that show that the way materials are presented influences importantly the way they are judged.

4.1.2 Experience. This section describes the results with a focus on the judges' experience with racing games. It is evident from Table 3 - p-value of 0.024 - that more experienced judges got better results compared to those with less experience or no experience at all. There are only 2 occasions where this does not happen. One of them is in the 'Single' type of questionnaire, where percentages are very similar. The other one is in those questionnaires with the shorter versions of the videos. In this second case, the accuracy of people with some experience is higher (albeit not significantly) than those with more experience, although these results need to be carefully considered given the fact that shorter videos generally produce less accuracy. It is also worth highlighting that irrespective of time and participant experience, longer videos achieved better results than shorter ones.

As for the perspective of the camera, overall, participants achieved better results with the game-like perspective over the TV perspective. The experienced players group had a slightly lower rate of correct answers, however this result exhibits quite a high standard error, which suggest that this might caused by an outlier. Finally, for the question parameter and as commented previously, only in experienced players 'Pair' was better than 'Single' and comparisons between videos could be affecting these results. In the final column, under 'All', we can see the percentage of the correct answers from all participants regardless of experience. This allows us to observe

Table 4: Accuracy of the judge evaluations to identify bots and humans, with standard error. Two ranges of confidence are explored: i) $0-4$ **for bots and** $6-10$ **for humans; and a more demanding ii)** $0-2$ **for bots and** $8-10$ **for humans**

Parameters	Type	Ranges	
		0-4 & 6-10	**0-2 & 8-10**
Question	**Single**	58.05% (2.27)	39.83% (2.25)
	Pair	55.15% (2.46)	32.60% (2.32)
Camera	**TV**	54.25% (2.49)	37.25% (2.42)
	Third	58.75% (2.25)	35.84% (2.19)
Time	**12 sec**	53.49% (2.69)	34.30% (2.56)
	35 sec	58.77% (2.13)	37.87% (2.10)

an overall result that matches the previous observations regarding the comparisons of the parameters.

4.1.3 Confidence. Table 4 shows results considering two different ranges for confidence in the evaluations performed by the judges. It is expected that correct answers with higher levels of certainty are less frequent than those in wider ranges, as results show - $p < 0.0001$. However, the difference between ranges is not overwhelming, with accuracies above 1 out of every 3 videos being labelled correctly. Most parameters seem to have maintained previous results - 'Single' questions with more certainty (perhaps given the lack of comparison and so, creating less doubt) and longer time, better accuracy. However, camera for lower threshold ranges was more accurate when using 'TV' rather than 'Third'. Perhaps, for some participants, given the unfamiliarity of the camera view, unusual behaviours may seem more suspicious and, hence, create more certainty upon seeing them, which would in turn produce higher rates in answers - which would itself increase the probability of getting the right answer.

4.1.4 Reasons for the evaluations. When it comes to the reasons behind participants' decisions, a wide range of answers were collected. The 'why' behind scores between 0 and 4 (which means they were judged as bots) ranged from how the agent planned its movement to how it interacted with others and, unfortunately, provided many contradictory answers: judges pointed out several times that they voted the agent as a bot because these held perfect timing in turns; were constantly centred behind other cars; too stable and uniform; far too perfect in turns; also predictable with a clear plan of path throughout the race and careful avoidance of crashes (for both car crash or leaving the track entirely from sliding/bad driving); a lack of flexibility in decision making; exaggerated speed and smoothness/fluidity which made it unnatural; and, unfortunately, they were wrong many of these times. However, participants also flagged agents as bots because, in their opinion, they acted too randomly with no reasoning behind decisions (with some pointing out that even a bad player would not attempt certain actions given that they were 'too pointless'); they thought the agents were 'good' at certain situations but not all (showing its inability of adapting); were far too extreme when accelerating or breaking; would not take on good opportunities when presented with them and would deliberately copy other agents' behaviours (which demonstrates a lack of learning skills and adaptation) with several pointing out the hesitation in action, delayed responses and no response to own surroundings and, once again, being wrong in their decision and were actually watching a human player. The only reasons that were given in a few occasions and were indeed correct were the feeling that it had 'robotic' and 'twitchy' movements and also adding the 'will' to crash on purpose.

Regrettably, reasons behind participants' decisions regarding what they believed to be videos with humans playing were no more enlightening. Some agents were classified as humans given their instability in driving - such as excessive steering/unsteadiness with a great degree of inconsistency as well which ranged from being too slow, doing unnecessary turns and having no notion of surroundings (not avoiding grass or water) to choosing the least optimal path, going off track and excessive crashing. Once again, in many occasions it was right but not all of them. However, other participants thought the agent they were judging was deemed to be a human because of its cautious nature, controlled speed and learning skills - by performing different actions in similar situations and adapting accordingly to new ones - with some commenting on their aggressive behaviour and boldness; ability of stopping an overtaking and hope of more crashes (by performing less perfect moves) - these were also reasons given to videos in which judges were sometimes right and sometimes were not.

This does demonstrate why the percentages of correct answers was never overwhelmingly correct or incorrect: there is always a particular reason that a judge will use to classify a bot as a human that another judge will use to classify a human as a bot. In addition, both the previous sets of answers seem to comply with one of the known limitations for believability assessment called self-deception [24] which is when something seems so real/unreal that the observant begins to rationalise his or her observations and believe in its truth (or lack of it).

In some occasions, some participants were also not sure about either of the options and would opt for a neutral response - 5, which means it could be either. In these cases, answers would most often be related to the possibility of the reasons given be applicable to both, as expected. This means that judges thought the driving style of the agent was good - which could mean it was an experienced player or just a good bot - or it was poor - from an inexperienced player or simply badly programmed AI - or just neutral - unimpressive but humans do make mistakes and AI could use simpler algorithms - with some expecting stranger behaviour from both.

It is also worth discussing some of the more detailed answers given throughout the study. Starting with some participants finding the 35 second videos not being long enough to judge but, at the same time, being pleased with the fact that they could rank the believability - given that in some occasions they thought they were observing a human for certain actions but not a perfect one still. Very few participants commented on comparing between videos when presented with two at the same time (Questionnaire type 'Pair') and also demonstrating some confusion when presented with the term 'believable'. We have also noticed that the more experienced players tended to use more technical terms and notice 'stranger' details such as 'speed of input' and the 'location' of the car when

racing - believing that the car was a bot/human because it was not in 'the middle of the road' and overall placement when trying to overtake and/or driving in a straight line. Some thought it was unfair for them to judge, precisely because they were very experienced in them and felt that they had discovered details that others would not; while others pointed out that it was hard for them to assess believability taking into account their lack of experience in racing games or unfamiliarity with the ones provided for this study.

Another interesting fact is the relevant amount of observations regarding judges' expectations. Many participants compared themselves to the agents they were watching and based their judgement on this. Some less experienced players thought that less skilled bots were humans because they reminded them of themselves when playing, and better bots were too good just like more experienced players expected certain behaviours, because they themselves tend to do it. All kinds of judges reported on how they expected AI to be 'worse' than the one showcased and humans better at certain actions with the opposite also being applicable - expecting AI to be better and humans worse. Furthermore, some situations apparently reminded participants of situations that they have been in and so, once again, they expected the agent to react as they would. Yet other interesting observations include the fact that some judges thought that the player was trying to fool them on purpose and, perhaps, the most important of these: the lack of interaction - some participants wished the videos were less simple and included more action for them to judge, especially between players. This comes as no surprise since one of the aspects that makes us humans is how we interact with each other. When we play video games we influence not only our own experience but also that of others. This also means that this interaction between agent and world could influence significantly how believable the agent is and the world itself - making this search for interaction logical for assessment in this study for some participants; given that people would have different cognitive capacities and some might prefer more time and/or more actions between agents/environment.

What the authors have concluded from these results is that the design of the questionnaire has great importance. Given one simple question, such as "How believable do you think this character is?", participants will interpret it differently - the concept of believability is not the same for everyone and a few did demonstrate some difficulty in which definition the authors were looking for. For this study, opinions were able to be grouped into more specific sections - there were always several reasons that were common in decisions for whether the agent was a bot or a human. Unfortunately, these reasons were not given to identify just NPCs or just players. Some participants would use one reason to flag an agent as a bot - such as "it cannot be a human it is 'too good"' - while other participants would use the exact same reason to flag an agent as a human - "it cannot be a bot it is 'too good"' - and, unfortunately, be right in some cases and wrong in others. This increases the difficulty in assessing believability as people will understand it in their own way - based on experience or background, etc.

4.2 Multiple Choice and Open Ended Question

As mentioned in Section 3.2.2, the two last questions involved asking participants whether they would prefer more believable characters in video games or not and if they wished to add any other information they found relevant. For the multiple choice question the majority (52) have voted to have more human-like characters in video games, followed by 'Maybe' (38) and 'No Opinion' (14) and, finally 'No' (6).

Answers to the open ended question mainly pointed out that the answer depends on the game and on the character's role - meaning that the agent should behave according to the situation it is in. It seems like for some games - such as those that involve some abstraction and emotions, like some single-player - it would be interesting to have agents that behave in a more human-like way; as interactions with characters would seem realistic and, perhaps, allow a more enjoyable experience. However, competitive games seem to also benefit from this characteristic; specifically those which provide tutorials and other modes to train players to face others in multi-player. Regardless, other games may not benefit from such - racing games and multi-player were given as examples. For the first, participants felt like it would be better to play against 'better AI' in order to increase their own skills. As for the latter, participants felt like it was rather deceitful substituting players with NPCs. It should be noted that not everyone decided to specify games or choose between 'yes' or 'no' but rather stay in a neutral position, where they do not hold a strong opinion in the matter but also have their own suspicions that certain situations would not be desirable; such as predicting human decisions.

Finally, the last topic discussed by a noticeable group of people was about suggestions for AI improvement. In these ideas, several proposed models to create something "more natural" such as introducing decision making into the programming (given the overuse of scripted behaviour), learning from mistakes - by analysing strategies that players execute, avoid them/apply them to others and/or alternate between these. This becomes fundamental in human-like behaviour as it is the very base of what distinguishes us from other agents in general. To include believable characters in games, learning, adapting and evolving becomes essential. These aspects should avoid extremely predictable actions and allow for new experiences and re-playability. Introducing believability is a mechanic that has to be thoroughly considered - as we have seen, it does not have to be applied in all settings and doing so might interfere with the enjoyment of a particular game.

5 CONCLUSION AND FUTURE WORK

Believability is a hard concept to grasp. The work described in this paper explores a version of the Turing Test adapted to a video game context. This study, however, targets the assessment methods and parameters rather than how human-like agents are playing. Several parameters were explored and introduced in our study and these include time, perspective, type of questionnaire, games and player experience.

Results show that changing parameters even if slightly can affect the overall results on believability assessment. Many reasons were also given for the choices in judgement which only showed how hard it is to access believability in general. Decisions seem to be very dependent on the game genre and how people play. Participants have judged many times based on how they play the games themselves and people have different playing styles - they will drive and react in

different ways which means it is worth identifying these types both in racing games and other genres in order to create player models. This would, in turn, allow for several 'versions' of believability, increasing the diversity of options and, perhaps, a more realistic setting. A majority of participants preferred more believable characters in video games but many commented on how game dependent this is. Many games seem to be appealing precisely for the fact that they have ridiculous behaviour implemented within their NPCs; which is not necessarily a bad thing. The point is to create games that are enjoyable to the people that play them. This only means that there is yet another possible research question taken from this: In which games should we apply human-like NPCs?

As mentioned, more than one game was selected to allow us results that could apply in a general setting within this genre. It is very likely that some games would take longer to assess than others and some would prove much harder than others to create an agent that can 'fool' a judge - racing games would be much less of a challenge than online games such as Massively Multiplayer Online (MMO) games which have a major interaction between players that includes speech and cooperation. This shows how important it is to apply these concepts in a more general setting. Finding aspects that make bots play like humans across several genres is, albeit a significantly difficult task, a major and relevant goal. We believe we are still far from achieving that, because even for a specific game there are many variables that may change drastically how the evaluation is performed (camera, judge involvement, judge experience, player experience, etc.). Moreover, even within the variables used, these are a representative but not an exclusive set. Other parameters that could also be explored, outside the scope of this study, are a a comparison between a binary judgement (Human/Bot, Believable/Not believable) and a ranged one $(0-10)$ as presented here; racing solo or competing with other players and/or NPCs; and even using more than two types per parameter (another range for session time, another camera perspective, etc.).

It is also worth mentioning that considering the different problems, such as self-deception and different capacities for evaluation, humans might not be the best for an accurate judgement but rather for helping modelling only. Another future possibility would be trying to collect only humans' and bots' data during play, turn them into 'signatures' and create an automatised method for assessing as proposed in [21]. This would reduce significantly human intervention and allow us to see how effective it is compared to the existing methods.

Thus, we can conclude that the first step is to try to identify the correct way of presenting the test to human (or bot) judges and, as shown in the present paper, that is still an open problem.

ACKNOWLEDGEMENTS

This work was supported by grant EP/L015846/1 for the Centre for Doctoral Training in Intelligent Games and Game Intelligence (IGGI) from the UK Engineering and Physical Sciences Research Council (EPSRC).

REFERENCES

[1] 2015. Drivatar and Machine Learning Racing Skills in the Forza Series, author=Jeffrey Schlimmer. nucl.ai. https://tinyurl.com/y9ogl99r

[2] Cyril Bossard, Romain Bénard, Pierre De Loor, Gilles Kermarrec, and Jacques Tisseau. 2009. An exploratory evaluation of virtual football player's believability. In *proceedings of 11th Virtual Reality International Conference*. 171–172.

[3] Lisa A Feldman. 1995. Valence Focus and Arousal Focus: Individual Differences in the Structure of Affective Experience. *Journal of personality and social psychology* 69, 1 (1995), 153.

[4] Bernard Gorman, Christian Thurau, Christian Bauckhage, and Mark Humphrys. 2006. Believability Testing and Bayesian Imitation in Interactive Computer Games. In *International Conference on Simulation of Adaptive Behavior*. Springer, 655–666.

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

[6] Philip Hingston. 2010. A New Design for a Turing Test for Bots. In *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*. IEEE, 345–350.

[7] Daniel Livingstone. 2006. Turing's Test and Believable AI in Games. *Computers in Entertainment* 4, 1 (2006), 6.

[8] Daniele Loiacono, Luigi Cardamone, and Pier Luca Lanzi. 2009. Simulated car racing championship 2009: Competition software manual. *Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy, Tech. Rep* (2009).

[9] Brian Mac Namee. 2004. Proactive Persistent Agents. *Using Situational Intelligence to Create Support Characters in Character-Centric Computer Games* (2004).

[10] Regan L Mandryk, Kori M Inkpen, and Thomas W Calvert. 2006. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology* 25, 2 (2006), 141–158.

[11] Jorge Muñoz, German Gutierrez, and Araceli Sanchis. 2010. A human-like TORCS controller for the Simulated Car Racing Championship. In *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*. IEEE, 473–480.

[12] Jorge Muñoz, German Gutierrez, and Araceli Sanchis. 2013. Towards Imitation of Human Driving Style in Car Racing Games. In *Believable Bots*. Springer, 289–313.

[13] Christopher Pedersen, Julian Togelius, and Georgios N Yannakakis. 2010. Modeling Player Experience for Content Creation. *IEEE Transactions on Computational Intelligence and AI in Games* 2, 1 (2010), 54–67.

[14] Diego Perez-Liebana, Jialin Liu, Ahmed Khalifa, Raluca D Gaina, Julian Togelius, and Simon M Lucas. 2018. General Video Game AI: a Multi-Track Framework for Evaluating Agents, Games and Content Generation Algorithms. *arXiv preprint arXiv:1802.10363* (2018).

[15] Mark O Riedl and R Michael Young. 2005. An Objective Character Believability Evaluation Procedure for Multi-Agent Story Generation Systems. In *International Workshop on Intelligent Virtual Agents*. Springer, 278–291.

[16] James A Russell. 2003. Core Affect and the Psychological Construction of Emotion. *Psychological review* 110, 1 (2003), 145.

[17] Edward F Schneider, Annie Lang, Mija Shin, and Samuel D Bradley. 2004. Death with a Story: How Story Impacts Emotional, Motivational, and Physiological Responses to First-Person Shooter Video Games. *Human communication research* 30, 3 (2004), 361–375.

[18] John R Searle. 1980. Minds, Brains, and Programs. *Behavioral and brain sciences* 3, 3 (1980), 417–424.

[19] Noor Shaker, Julian Togelius, Georgios N Yannakakis, Likith Poovanna, Vinay S Ethiraj, Stefan J Johansson, Robert G Reynolds, Leonard K Heether, Tom Schumann, and Marcus Gallagher. 2013. The Turing Test Track of the 2012 Mario AI Championship: Entries and Evaluation. In *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*. IEEE, 1–8.

[20] Noor Shaker, Julian Togelius, Georgios N Yannakakis, Ben Weber, Tomoyuki Shimizu, Tomonori Hashiyama, Nathan Sorenson, Philippe Pasquier, Peter Mawhorter, Glen Takahashi, et al. 2011. The 2010 Mario AI Championship: Level Generation Track. *IEEE Transactions on Computational Intelligence and AI in Games* 3, 4 (2011), 332–347.

[21] Fabien Tencé and Cédric Buche. 2010. Automatable evaluation method oriented toward behaviour believability for video games. *arXiv preprint arXiv:1009.0501* (2010).

[22] Fabien Tencé, Cédric Buche, Pierre De Loor, and Olivier Marc. 2010. The challenge of believability in video games: Definitions, agents models and imitation learning. *arXiv preprint arXiv:1009.0451* (2010).

[23] Frank Thomas, Ollie Johnston, and Walton Rawls. 1981. *Disney Animation: The Illusion of Life*. Vol. 4. Abbeville Press New York.

[24] Julian Togelius, Georgios N Yannakakis, Sergey Karakovskiy, and Noor Shaker. 2013. Assessing Believability. In *Believable bots*. Springer, 215–230.

[25] Alan M Turing. 2009. Computing Machinery and Intelligence. In *Parsing the Turing Test*. Springer, 23–65.

[26] Niels Van Hoorn, Julian Togelius, Daan Wierstra, and Jurgen Schmidhuber. 2009. Robust player imitation using multiobjective evolution. In *Evolutionary Computation. IEEE Congress on*. IEEE, 652–659.

[27] David Weibel, Bartholomäus Wissmath, Stephan Habegger, Yves Steiner, and Rudolf Groner. 2008. An Evaluation of Perceived Personality in Fictional Characters Generated by Affective Simulation. *Computers in human behavior* 24, 5 (2008), 2274–2291.

[28] Georgios N Yannakakis, John Hallam, and Henrik Hautop Lund. 2008. Entertainment capture through heart rate activity in physical interactive playgrounds. *User Modeling and User-Adapted Interaction* 18, 1-2 (2008), 207–243.