frontiers in Computer Science

ORIGINAL RESEARCH
published: 23 December 2021
doi: 10.3389/fcomp.2021.774763

Assessing the Believability of Computer Players in Video Games: A New Protocol and Computer Tool

Cindy Even[1], *Anne-Gwenn Bosser*[2] *and Cédric Buche*[3]*

[1]Virtualys, Brest, France, [2]Lab-STICC, CNRS, ENIB, Brest, France, [3]IRL CROSSING, CNRS, ENIB, Adelaide, SA, Australia

In this paper, we address the challenge of believability in multiplayer video games. Our contribution is a system for assessing the believability of computer players. The state of the art examines existing methods and identifies seven distinguishing features that differ considerably from one assessment to the next. Our investigation reveals that assessment procedures typically alter gameplay, posing a considerable danger of bias. This is a major flaw since computer players are evaluated in a specific context rather than in the context of the game as it should be played, potentially skewing the findings of the evaluation. As a result, we begin on a trial-and-error process, with each new proposal building on the achievements of the previous one while removing the flaws. New proposals are tested with new assessments, a total of three experiments are then presented. We created a computer program that partially automates the execution of the assessment procedure, making these trials easier to implement. At the end, thanks to our proposal, gamers can assess the believability of computer players indirectly by employing reporting forms that alert users to the presence of bots. We assume that the more a bot is reported, the less credible it becomes. We ran a final experiment to test our proposal, which yielded extremely encouraging results.

Keywords: human-centered, interaction design, virtual player, bot, video game, assessment

OPEN ACCESS

Edited by:
Kenny Mitchell,
Edinburgh Napier University,
United Kingdom

Reviewed by:
Ricardo José Vieira Baptista,
University of Porto, Portugal
Claudia Krogmeier,
Purdue University, United States

***Correspondence:**
Cédric Buche
cedric.buche@cnrs.fr

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *13 September 2021*
Accepted: *18 November 2021*
Published: *23 December 2021*

Citation:
Even C, Bosser A-G and Buche C (2021) Assessing the Believability of Computer Players in Video Games: A New Protocol and Computer Tool. Front. Comput. Sci. 3:774763. doi: 10.3389/fcomp.2021.774763

1 INTRODUCTION

The popularity of a video game can be greatly influenced by the implementation quality of its computer players (Scott, 2002). Modern multiplayer video games do not require unbeatable bot, the goal is now to get believable behaviour (Livingstone, 2006; Soni and Hingston, 2008). Different approaches have been adopted for the development of believable bots, such as systems based on connectionist models (Van Hoorn et al., 2009; Llargues Asensio et al., 2014), production systems (Laird and Duchi, 2001; Polceanu, 2013) and probabilistic models (Le Hy et al., 2004; Gorman et al., 2006; Tencé et al., 2013), to mention just a few.

Generally, the proposed systems are not assessed. When they are, the results can not be compared as different protocols have been used. However, many authors (Mac Namee, 2004; McGlinchey and Livingstone, 2004; Gorman et al., 2006) pointed out the need of a generic and rigorous evaluation that would allow the comparison of new systems against existing ones. The evaluation of Artificial Intelligence (AI) in games research has been identified as one of the main challenges in game AI research (Lucas et al., 2012). Although the evaluation of bots' performance can be performed through objective measures, the evaluation of bots' believability is complex due to its subjective aspect.

The objective of this article is to provide a solution for assessing the believability of computer players in multiplayer video games. **Section 2** provides a literature review of the protocols previously used to assess the believability of computer players. After analysing, we identified seven features that characterise the assessments and which vary significantly from one to another. When designing a new protocol, these features need to be chosen carefully in order to not introduce a bias into the evaluation. After an in-depth analysis of these protocols, we give recommendations for the features that are well established. We also identify the other features that still need further study and testing to be determined. To facilitate the execution of the evaluation, in **section 3** we developed a system that partially automates the evaluation process. Its structure and implementation are also presented in detail. In **section 4** we present our first protocol proposal. During the literature review we found out that the video game's gameplay could be affected by the assessment process. To avoid this we sought to hide the purpose of the evaluation by building a questionnaire aiming attention at several aspects of the game, the goal being to disperse the attention of the participants on the whole game rather than simply on their opponent. **Section 5** presents the evaluation that we had the chance to organise during a competition that took place at the national conference. We took advantage of this event to profile the judges according to their ability to correctly distinguish bots from human players. The method used to carry out this experiment as well as the results obtained are provided in details. A preliminary version of this section has been reported in (Even et al., 2018), lacking from connection with the following proposition. From the observations that we could make during our previous experiments, we came up with a completely new design, detailed in **section 6**. For this new approach we tried to use the game as it is normally played, with the aim of minimising as much as possible the impact of the assessment on the gameplay. We decided to take inspiration from the reporting systems already present in many video games. Once again we describe the experiment we carried out to evaluate our approach and present the promising results we obtained. In **section 7** we conclude with a summary of the work we have done and we provide some prospects for improvements.

2 STATE OF THE ART

According to Bates (1994), the notion of believable characters refers to the illusion of life and permits the audience's suspension of disbelief (Coleridge, 1817). To create this illusion, animators defined references (Thomas and Johnston, 1981). In video games, users can interact with their characters and inhabitants. The notions of presence (Schuemie et al., 2001) and co-presence (Goffman, 1963) or social presence (Heeter, 1992) are often evaluated and can be measured using self-report or behavioural measures (Bailenson et al., 2004). Agent's believabilty is frequently adressed (Magnenat-Thalmann et al., 2005; Bosse and Zwanenburg, 2009; Bevacqua et al., 2014), determined by aspects such as emotions, personality, culture, style, adaptation to the context, and many others (Poggi et al., 2005).

The concept of believability for characters in video games can be divided into two broad classes (Togelius et al., 2012): character believability and player believability. Character believability (Loyall, 1997; Bogdanovych et al., 2016; Verhagen et al., 2013) refers to the belief that a character is real. In this case, the notion of believability coincides with the definition in character arts and animation. On the other hand, player believability refers to the belief that a character is controlled by a human player (Tencé et al., 2010). Its behaviours are the result of some ongoing input from a human player who is aware of what the character is doing in the game. The next section offers a review of existing methods used to assess believability of computer players.

2.1 Assessing Believability

2.1.1 Competitions

In recent years we have seen the emergence of competitions (Togelius, 2016) oriented toward the implementation of human-like (or believable) opponents such as the 2K Botprize competition (Hingston, 2009). It is a variant of the Turing test (Turing, 1950) which uses the "Deathmatch" game-type mode of the video game. The objective is to kill as many opponents as possible in a given time (and to be killed as few times as possible). In its first two editions (Hingston, 2009) the protocol includes five rounds of ten minutes. In each round, each human judge was matched against a human confederate and a bot. The confederates were all instructed to play the game as they normally would. At the end of each round, the judges were asked to evaluate the two opponents on a rating scale, and to record their observations. In order to pass the test, a candidate (a bot or a human player) was required to be rated 5 (this player is human) from four of the five judges. In 2010, a new design was implemented to make the judging process part of the game. A weapon of the game could be used to tag an opponent as being human or bot. If the judgement was correct, the result was the death of the target, if incorrect, the death of the judge's avatar. Both bots and humans were equipped with the judging gun and could vote. This modification to the system introduced a bias in the evaluation process as the gameplay was adversely affected (Thawonmas et al., 2011; Polceanu, 2013). The only differences were that the judges would not die if they made a wrong judgement. They could change their judgement by tagging the candidate again. Only the tag in place at the end of the game was taken in account. With this new rule, the judges would not know instantly if they had made a mistake or not, which would stop them from changing their judging strategy and would make them judge every candidates the same way. In 2014, the novelty was the addition of a third-person believability assessment (i.e., the judges observe the game). While performing the former method, the matches were recorded on the server. Clips were then selected from these videos and used with a crowd-sourcing platform where users could vote after watching each clip. Different opinions emerge when it comes to chose whether the assessment should be first

TABLE 1 | Comparison of the existing experiments from (Even et al., 2017).

Reference	Application	1st or 3rd person assessment		Duration	No. of judges	Judges' level			Information given	Subjective assessment type				How
						Novice	Medium	Expert		Binary	Comparison	Scale	Comments	
Laird and Duchi, (2001)	Quake II Deathmatch	3rd	16 × 1 video candidate's view	3 min	8	✓		✓	A	✓		✓ 1 to 10		n/a
Mac Namee, (2004)	Simulation of a bar	3rd	2 simulations global view	as long as needed	13	✓	✓		B		✓ 2 choices	✓ 1 to 5		pen and paper
McGlinchey and Livingstone, (2004)	Pong game	3rd	video global view	n/a	n/a	n/a	A		✓ 4 choices		✓	n/a		
Gorman et al. (2006)	Quake II Deathmatch	3rd	15 × 3 videos 1st person view	20 s	20	✓	✓	✓	A			✓ 1 to 5	✓	online
Bossard et al. (2009)	CoPeFoot	1st		n/a	48	✓		✓	C	✓				pen and paper
Hingston, (2009) (BotPrize v1)	UT2004 Deathmatch	1st		10 min	5		✓	✓	A			✓ 1 to 5	✓	n/a
Hingston, (2010) (BotPrize v2)	UT2004 Deathmatch	1st		n/a	7		✓	✓	A	✓				in-game
Llargues Asensio et al. (2014) (BotPrize v3)	UT2004 Deathmatch	1st		15 min	3		✓		n/a	✓				in-game
		3rd	10 × 1 video 3rd person view	1 min	12		✓		n/a	✓				Crowdsourcing platform
Acampora et al. (2012)	UT2004 Capture The Flag	3rd	1 × 4 videos 1st person view	n/a	10		n/a		n/a			✓ 1 to 7		n/a
Shaker et al. (2013)	Infinite Mario Bros	3rd	2 videos global view	1 min	73		n/a		n/a		✓ 4 choices			online
Bogdanovych et al. (2016)	Everyday life of the Darug people	3rd	14 × 2 videos 1st person view	n/a	43		n/a		B		✓ 3 choices	✓ 1 to 5		online

Character believability assessment highlighted in red.
A Judges are told that there is a mix of bots and humans.
B Judges know the nature of each entity.
C Judges are given no information.

or third person oriented or a combination of both as it has been done here.

In Acampora et al. (2012), the objective of the game is to capture the enemies' flag and to return it to your own team's flag. The authors suggested two ways to assess the believability of bots. The first one used objective measures: the score of the bot and the duration of the match. The authors made the assumption that a believable bot should have a medium score and that the duration of the match should be relatively high. However, the results obtained can be questionable as the most believable bot was the one with the lowest score and who played the longest match. Their second assessment used subjective measures: 20 videos were recorded where an expert player played against bots and human players with different levels (novice, medium, and high). After watching four videos, judges were asked to evaluate human-likeness on a 7-point Likert scale. Laird and Duchi (2001) and Gorman et al. (2006) used similar approach. The protocol's characteristics of these player believability assessments can be found in **Table 1** along with other relevant references of player believability (in white) and character believability (in red) assessments.

2.1.2 Criteria-Based Assessment

Some authors have worked on criteria-based assessment methods where the believability of bots is ranked by the amount of criteria they meet. Hinkkanen et al. (2008) proposed a framework that is composed of two aspects: firstly, character movement and animation, secondly, behaviour. Each criterion is worth a certain number of points depending on its impact on credibility. Each time a bot fullfils one of the criteria, it gets the points that are then added up to get a score. An overall score is obtained by multiplying the scores from both aspects. A much more detailed solution was offered with ConsScale (Arrabales et al., 2010). This scale is directly inspired by an evolutionary perspective of the development of consciousness in biological organisms. It aims at characterizing and measuring the level of cognitive development in artificial agents. A particular instantiation has been performed for First Person Shooter (FPS) game bots (Arrabales et al., 2012), specifying a hierarchical list of behavioural patterns required for believable bots. This list consists of 48 cognitive skills spread over 10 levels. However, judges from the 2010 edition of the BotPrize reported that the list is interesting and appropriate but that it is difficult to take all the subtler points of the scale into account during the assessment. Such solutions are rather intended to provide a roadmap for the design of human-like bots. They allow to show the presence or absence of some specific features that could have an impact on the final result.

2.2 Analysis

The protocols used in the past for the assessment of computer player's believability have characteristics that vary significantly. The process of judging the behaviours of a bot is by nature a subjective process (Mac Namee, 2004; McGlinchey and Livingstone, 2004; Livingstone, 2006) as it depends on the perceptions of the people playing or watching the game. Having no obvious physical attributes or features that can be measured, the only solution for measuring the believability of bots that can be considered is the use of a questionnaire (Mac Namee, 2004). In some cases, the players fill the questionnaire after playing the game for some minutes, in other cases they vote during the game. The judgement can be done by the players or by observers, and different types of questionnaires are used such as ranking or comparison. In this section we propose to analyse characteristics of the protocols collected in **Table 1**. When studying the protocols used in the past to assess computer players' believability, we identified some characteristics that varied significantly from one assessment to another, giving results that can not be correlated.

First of all, different types of games were used such as FPS, sport or platform games. The main criterion when choosing the game is that it needs to be a multi-player game where one can face computer players. The second criterion, which restricts significantly the range of games that can be considered, is that it has to be possible to interface a bot.

Even when the types of games used in the assessments were similar, judges had different roles. They were either part of the game (first person assessment), with the ability to interact with the candidates but also with the risk of modifying the gameplay. Or they were spectators (third person assessment), assessing a game in which they were not involved. For this type of assessment, the judges watch videos of the game. These videos can be recorded using different points of view. The most commonly used is the confederate's first person view but a solution that seems to have potential and needs to be tested is the candidate's third person point of view.

The duration of the assessment is another characteristic that can vary significantly. Judges might give a random answer if they do not have enough time to evaluate a bot. In order to avoid this situation it seems important to define a minimum assessment duration.

As the notion of believability is very subjective, it is important to collect a large number of judgements. The use of an on-line questionnaire or crowd-sourcing platform seems unavoidable as they can allow for the collection of more data that would give more accurate results. In order for the protocol to be rigorous, a minimum number of participants must be defined.

The judges' and confederates' level of experience is sometimes taken into account. In general, we recommend training novices before involving them in the roles of judge or confederate as they need to know the rules, the commands and to have experimented with the game. Otherwise, confederates could easily be mistaken with weak bots and judges could be too confused to be able to make a judgement. It would be interesting to study the influence of the judges' level on the results when the number of judges is high.

Finally, different types of questionnaire have been used (binary, scale or comparison) to collect the judges' opinions, giving data that can not be compared from one assessment to another. Regardless of the type of questionnaire, the question(s) as well as the offered solutions will have to be adapted according to the type of assessment (first or third person) and the information previously given to the judges.

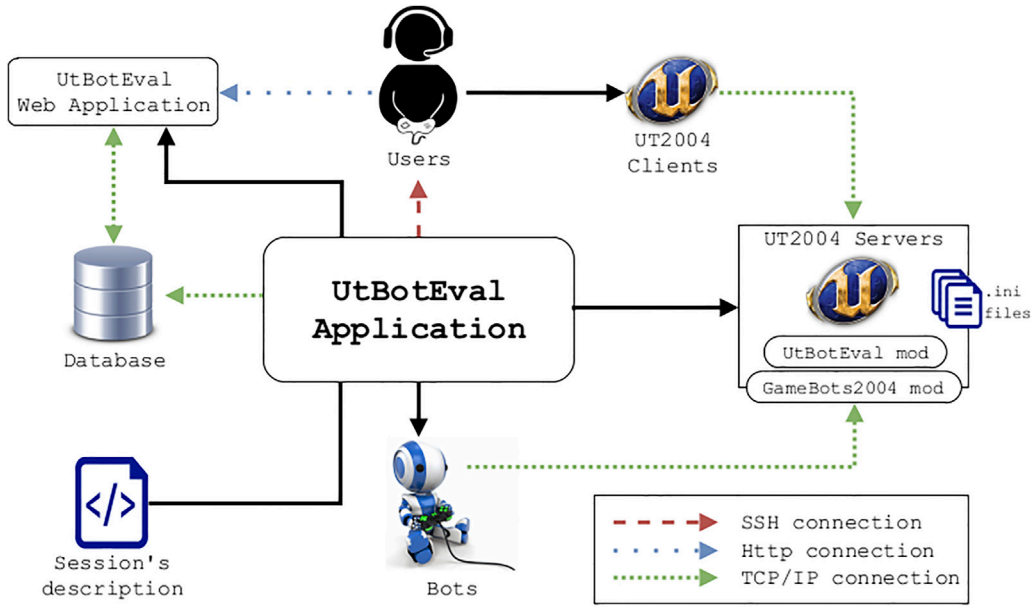

FIGURE 1 | System architecture of the UtBotEval system.

2.3 Conclusion

In this section we analysed the protocols previously used to assess the believability of computer players. We identified seven features that characterise the assessments and which vary significantly from one to another. When designing a new protocol, these features need to be chosen carefully in order to not introduce a bias into the evaluation. The design of a new protocol should be easy and flexible to test various configurations. From an implementation point of view, current solutions does not offer such an operational computer tool.

3 UTBOTEVAL: A TOOL FOR ASSESSING THE BELIEVABILITY OF BOTS

3.1 UtBotEval

To facilitate the evaluation process and to prevent the judges from performing additional manoeuvres such as connecting to a specific server to start playing the game, we developed a system that partially automates the assessment. It is composed of three specific modules linked together via various communication protocols (see **Figure 1**). We used the same video game as for the BotPrize competition: UT 2004. However, other games can be used as long as it is possible to run a dedicated server and to connect players and external computer programs (bots) to it.

3.2 UtBotEval Application

The UtBotEval application is the core of our system. Its structure is described with the UML class diagram in **Figure 2**. The main class of our framework is the Procedure class, it is a singleton whose role is to control the progression of the experiment from beginning to end. It is composed of a list of players containing an instance of Human for each participant as well as an instance of Bot for each bot to evaluate. The Procedure can start the web server (WebServer.Start ()) and remotely (via SSH) open a web page displaying instructions or explanations on the video game for example, with the Human.RunCmd() method.

We structured the course of the matches so as to facilitate their management. Thus, each group of participants takes part in a Session. It is composed of several rounds::Round[*] being themselves made up of several matches::Match[*]. Each match requires a dedicated game server (UtServer) on which two players face each other. Take for example a situation where four participants have to evaluate three bots. They will take part in a new session, consisting of one round of training and six rounds of evaluation (one for each bot plus one for each opposing human). Each round will have from two (two matches of human against human) to four matches (each human confronts a bot). For each session, the order in which the participants will meet their opponent as well as the name of the map to be used for the game is given in a descriptive file (xml or json). The Session.GetRounds() method allows to instantiate the matches and rounds by respecting this specific order.

The UtServer class is used to manage the dedicated servers of UT2004. Several parameters can be entered when starting a new game server such as the name of the map, the maximum duration of the match (TimeLimit), the maximum score (GoalScore), the mod and the.ini file. The mod can either be a native game type such as the Deathmatch (which is used for the training phase), or a custom mod such as the UtBotEval mod (described bellow). To facilitate the organisation of the evaluation we decided to run all the servers on a single computer. However, this implies that the servers have different IP addresses to be able to choose which one to connect to. To do this, each server must have its own.ini file where is specified its assigned port number. This works for

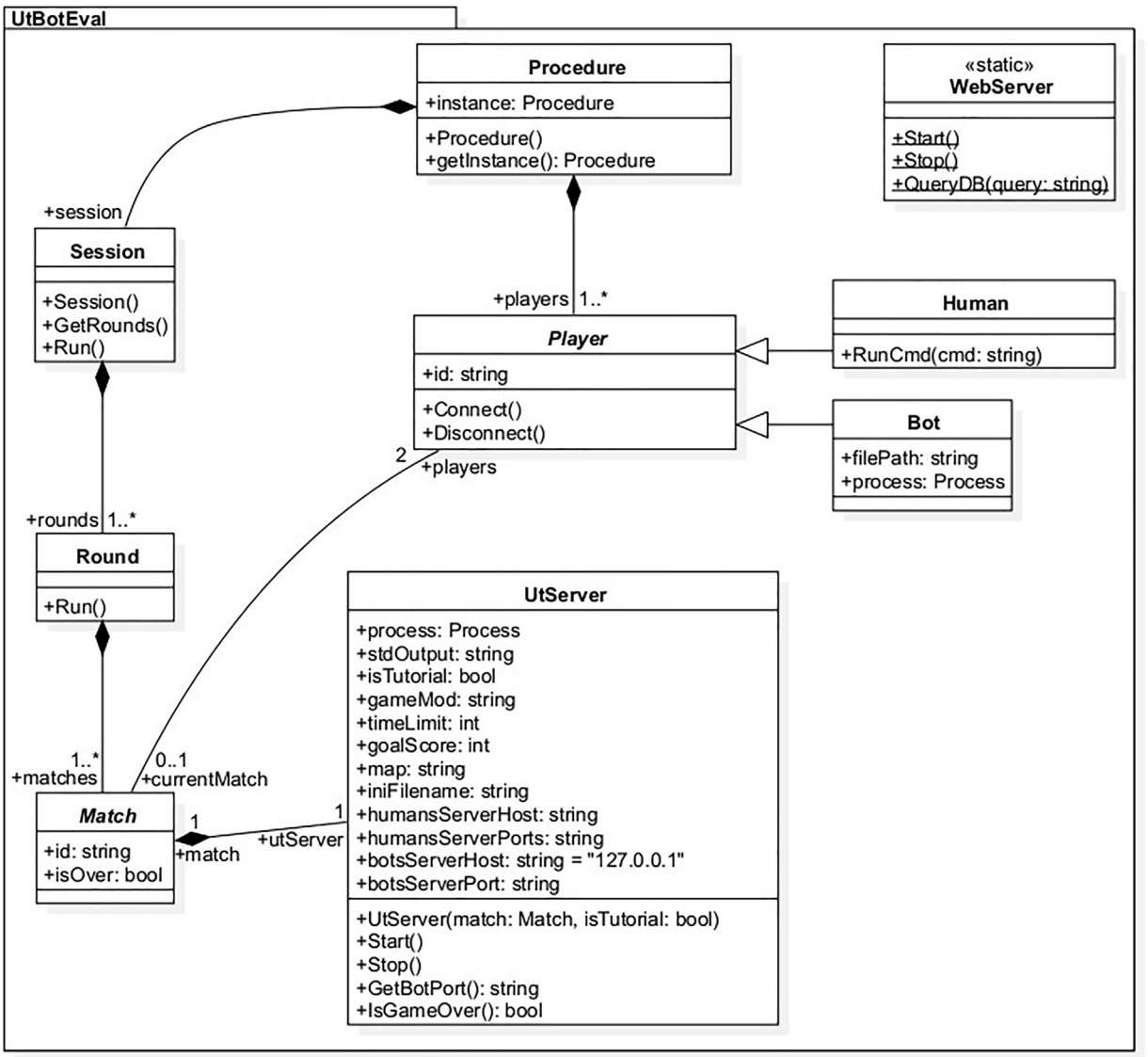

FIGURE 2 | UML class diagram of the UtBotEval Framework.

human players only but the problem is the same for bots. As we mentioned earlier, bots can connect through the GameBots2004 mod. By setting the value of bRandomPorts to True in the GameBots2004.ini file, each server uses a random port number for the connection of bots. The UtServer.GetBotPort() method can retrieve this port and update the botsServerPort value. The UtServer.IsGameOver() method checks at regular intervals if the match is over. If this is the case then the method returns True and the evaluation can continue.

Matches are started with the method Match.Start() which automatically starts the utServer and connects the players to it. The methods Player.Connect() and Player.Disconnect() are abstract since their implementation depends on the type of player. Human players are remotely connected to the game server with an SSH command ordering the opening a new game client window with the server IP address in parameter. Bot players on the other hand run on the same computer as the game servers. The connection consists in starting a new process with the game server IP address in parameter.

When all the matches from a round are finished, bots, game clients and game servers are automatically stopped. Participants are then directed to a web page displaying the questionnaire. Once they have finished giving their answers, the round is over and the next one of the session can begin.

3.3 UtBotEval Mod for UT2004

UT2004 includes extensive modification support which allows users to easily create maps, models and game modes, as well as various other additions to the game. A mod was developed specifically for the evaluation. It has a class that inherits from

the BotDeathMatch class of GameBots2004. This allows us to make changes when bots and players join a DeathMatch game server (see the code bellow). In the game, players are represented by their avatar in the 3D environment and the player's name is displayed above this avatar. To make sure that the participants do not have a clue about the nature of their opponent from their name or appearance, our mod provides anonymity to the players in a similar way that the BotPrize mod does thanks to the methods getCharacter and ChangeCharacter. When a player (human or bot) connects to the server (with the methods AddRemoteBot, AddEpicBot and Login), he is assigned a name and a skin (the player's appearance) which are randomly selected from the list of the default players in the game (provided in the defaultproperties). Access to the chat, scoreboard and players' statistics have also been removed in the users' settings to prevent the participants from accessing meta-gaming information that could help them to distinguish between bot and human.

4 EXPERIMENT 1: BLINDING THE JUDGES

(MacLean and Dror, 2016) (Gilovich et al., 2002; Koehler and Harvey, 2004). To date, the gameplay of the game is affected by the state-of the art evaluation process. To reduce the risk of bias in scientific experiments, "blinding" techniques can be used (MacLean and Dror, 2016). Therefore, participants were blinded to the objective of this experiment. To achieve this, we built a questionnaire in such a way that the main question was hidden among others. By adding many questions that deal with different aspects of the game we hoped to disperse the participants' attention on the whole game rather than on a specific item: the opponent. The assessment runs in a number of rounds, similar to the format of the first version of the BotPrize competition (Hingston, 2009). However, some changes have been made. To avoid revealing the purpose of the experiment, participants are simply informed that the they would take part in an experiment about video games. To allow a more in depth assessment without the distraction of a third player, we made the choice to only play one-on-one matches. Confederates are no longer necessary, instead, in each match a judge will play against a bot or against another judge (Hingston, 2009); (Yannakakis and Martínez, 2015); (Dolnicar et al., 2011); Krosnick (2002); Hingston (2009).

4.1 Method

4.1.1 Participants

Four groups of four students (16 participants) an engineering school participated in the experiment. The participants were all volunteers and no compensation was provided for their engagement.

4.1.2 Procedure

Participants are recruited in groups of four and are only informed that the experiment is about video games. They are provided with the following indications.[1]:

"This experiment lasts approximately 1 hour and uses the video game: Unreal Tournament 2004. It will begin with a training phase. After quickly reading the rules of the game, the participant will play a training match. Then a questionnaire will appear, the first time we do not take into account the answer since it is the training phase. Then, the participant will play several matches of the game. Each match will have a different configuration. At the end of each match, the participant will have to quickly fill the questionnaire evaluating his feeling towards these different configurations. The participant will have to concentrate on the objective of the game: to kill a maximum of times his opponent while being killed a minimum of times. Finally, a last questionnaire will be provided at the end of the experiment."

The protocol of the original BotPrize was adapted to our needs. We kept the presentation similar with Hingston (2009) to facilitate the comparison:

A) Training phase. In order to familiarise the participants with the game, training phase consists of providing information about the game, its controls, weapons and power-ups. Then, participants play a 3-min game against a native bot of the game. Finally, the questionnaire is displayed which ensures that the participants will be in the same conditions for the evaluation of all its opponents.
B) For each judging round:
   1) The servers were started.
   2) When the matches involved bots, they were started and connected to their assigned server.
   3) The judges were automatically connected to the game on their assigned server.
   4) Each game was a Death Match.
   5) At the end of the round, each judge was asked to fill the questionnaire.
   6) After a short break, the next round starts.
C) Final questionnaire.

For our experience we decided to evaluate five bots. Thus, each participant played eight games facing the five bots and three other participants one after another. All participants must encounter all their opponents on a different map. The order we used for the experiment was generated partially randomly to meet this constraint.

4.1.3 Measures

We made the assumption that by adding questions about different aspects of the video game in the questionnaire, the real purpose of the experiment (the evaluation of the bots) would be hidden. We will therefore check whether this objective was unmasked or not by the participants. We developed a questionnaire with several themes to avoid the judges focusing only on their opponent, which would have the effect of changing the gameplay. The original version is composed of three questions about music, two about the opponent, one about the duration of the match and four about the map. Among these categories, different types of questions are used: three questions ask for the participant's feeling, and seven questions require a degree of

[1]Translated from French.

certainty (four of which have three possible answers and three have only two choices). For the question used to evaluate opponents' believability, rather than using a five-level Likert scale like for the BotPrize (Hingston, 2009), we used a binary scale coupled with a certainty scale. While previous work (Yannakakis and Martinez, 2015) encourages the use of rank-based questionnaire over rating-based questionnaires, we could not use this method as it only applies to situations where participants are asked to rank two or more players. Thus, we decided to use a binary scale. This type of scale has been proven to be equally reliable, quicker and perceived as less complex (Dolnicar et al., 2011) than traditional rating-based questionnaires. In case the participant hesitate between two proposals, we have added the possibility for them to give their degree of certainty. Some may argue that a simple "I do not know" option would have been sufficient. However, according to Krosnick (2002), adding this option can result in the decision not to do the cognitive work necessary to give a proper response. To avoid this, we forced the participants to:

1) choose between A and B: I believe that the opponent was controlled by (A) a computer program, (B) a human;
2) give their degree of certainty on a ten-level scale going from "Not sure at all" to "Completely sure".

A questionnaire was added at the end of the experiment to allow us to verify if the objective was not discovered by the participants. It is composed of four questions:

- What do you think the purpose of the experiment was?
- At what point (approximately) did you understand the objective? First round/second/[. . .]/eighth.
- Did you change the way you played the game? Yes/No.
- Do you have any remarks.

This questionnaire is simply intended to evaluate our approach and should not be present when using this protocol to assess bots' believability.

4.2 Results

We analysed the answers given in the final questionnaire. Unfortunately the results were not as expected. Out of sixteen participants, eleven (61%) discovered the real objective of the experiment. We considered that the objective was discovered when the participant mentioned the evaluation of the opponent/artificial intelligence/bot in his/her commentary. The second question did not satisfy us either, as five of the participants (28%) said that they had changed their way of playing for the experiment. We concluded that these two results were too high and therefore the experiment did not meet the objective we had set.

4.3 Discussion

Our method of keeping the goal of the experiment secret was clearly a failure. Indeed, more than half of the participants have guessed the objective and more than a quarter felt they had changed the way they play the game, which is exactly what we wanted to avoid. In order to improve our proposal, it would be interesting to get more information about the participants and especially about their expertise in video games. In fact, we believe that students of school of engineering are particularly familiar with video games as they are trained in computer programming during their studies and that video games are regularly used in practical work. This familiarity could be the reason why they gave a particular importance to the artificial intelligence of their opponent rather than other aspects of the game present in the questionnaire like the music and the map of the level. We will investigate these elements in the next section.

However, our technical setup has proven to be very efficient and easy to use. It allows the investigator not to worry about starting the game servers by hand and connecting the right players to it. This allows to avoid any mishandling that could disrupt the progress of the experiment. This system is also very flexible because it allows us to easily manage the number of participants and bots that we want to include in the evaluation as well as the duration of the games and the number of matches that participants must play. The same system architecture was used for the other two proposals we have made, which are presented in the two sections to follow.

5 EXPERIMENT 2: INFLUENCE OF THE JUDGES' EXPERTISE

In the previous experiment, our attempt to mask to the judges failed. In this experiment, we hypothesis this was due to the participants' familiarity with video games. Therefore we study the impact that the level of expertise of judges in video games could have on their ability to distinguish bots from human players (Even et al., 2018).

5.1 Method

5.1.1 Participants

Six teams competed in the tournament (one team == one bot), with three making it to the finals. 60 members of the national artificial intelligence research community took part as jury. In the rest of the section, the term "participants" refers to individuals who participated in the jury and not the competitors as they were not present during the competition.

5.1.2 Material

In this new experiment, we do not intend to try to avoid gameplay modifications (Even et al., 2018). Therefore, we simplified the questionnaire of our previous protocol to keep only the question used to judge the opponent. As a result, we urge participants to:

1) choose A or B: I think the opponent was controlled by (A) a person or (B) a computer program;
2) give level of certainty on a ten-point scale ranging from "Not sure at all" to "Completely sure".

We also modified the final questionnaire to collect information on the judges' playing habits. Since we did not find an existing questionnaire to estimate the gamers' level in

video games, we built our own with questions deemed relevant to estimate their familiarity with the type of game used for the assessment and with the presence of bots.

We reused the computer tool described in **section 3** to save time and ensure a large number of participants. The only changes we had to make to the system were an update to the questionnaire in the web application and the addition of a new ending condition to our UT2004 mod. We also utilised our mod to record information about the contest, such as the duration and player scores.

We organized a competition, the BotContest.[2], as part of the French Association of Artificial Intelligence conference. The aim was obviously to develop the most believable bot for the video game Unreal Tournament 2004. During the finals we had the opportunity to use our protocol and obtain useful information regarding the judges and their level of expertise in video games.

Thanks to feedback from the previous experiment, we apply the following modifications. First, we set a maximum duration to make sure all the participants would play the same amount of time with all their opponents. However, we observed that the number of times the competitors encountered their opponent varied dramatically from one match to the next. Similarly, when establishing a score to win the match ("GoalScore"), the number of times the players meet might range from n (i.e., one player gets all the points) to $(2n - 1)$ (i.e., the game is tied until the last shot). We modified the GoalScore's behaviour to make the competition more equitable. We chose to keep track of the overall number of frags.[3] that occurred throughout the contest. When a frag happens in the game, a counter is incremented, and when this number reaches the limit defined by the GoalScore parameter, the game is immediately terminated. Suicides are not included in the frags since they are rarely caused by the opponent. Suicides can take several forms, including plunging into a pit, lava, or acid, shooting yourself, or being killed by your own weapon's discharge. We also use a TimeLimit as a safety net to ensure that the game does not go too long due to logistical constraints. To avoid confusion with the game's original GoalScore parameter, we'll refer to this parameter as "FragLimit" in the rest of the paper.

5.1.3 Procedure

When the participants arrived, a web page with the following instructions was already open.[4]:

"Here is your mission, you will have to play against several players one after the other. These players might be controlled by one of the programs sent to us for the competition, or by another human player. After each game, you will have to fill a form to say if you think your opponent was controlled by a human or computer program. You will also need to specify your degree of certainty. For example, if you are unable to tell if your opponent is a human or a bot, you can check a response (bot/human) randomly and put the cursor on "Not sure at all". During games, it is important that you play the game as you normally would, do not change the way you play because of the judgement. When you are ready to start, click on the "Continue" button".

After then, the experiment continued with a training phase. The experiment's second phase consisted of four rounds in which participants played a game of UT2004 with the BotContest mod and then filled out the judgement form after each game. The contestants would face the three bots and one of the other participants over the four rounds. This information was obviously kept hidden from the participants, who only knew that they would be pitted against a random number of bots and humans in a random sequence. Participants were asked to complete a questionnaire that collected personal information about their gaming habits in the final phase (see **section 5.1.5** for a detailed description.).

5.1.4 Variables

We used four distinct "maps" from the game for this experiment: DM-1on1-Albatross, DM-1on1-Spirit, DM-1on1-Idoma, and DM-Gael. We chose these maps because of their tiny size, which is ideal for one-on-one deathmatch battles. The DM-Gael map, for example, was picked for its unique feature of having only one major chamber with a rather wide and deep hole in the centre. A platform floats in the middle of the pit, where power-ups can spawn. Reaching this pickup is dangerous, since falling down the pit will kill you. "TimeLimit" was adjusted at 5 min, making the entire experiment last around 30 min, in order to accommodate hosting conference limits and a threshold discovered during the preliminary qualifying procedure. During the qualifying process, it was discovered that certain bots could not sustain credible behaviour over time. Some began to exhibit repetitive and predictable behaviour after 3 minutes, such as going back and forth or always taking the same path or employing the same assault technique. As a result, we concluded that a match's duration should be higher than 3 minutes. After considerable testing, "FragLimit" was set to a value of ten. We previously utilised a FragLimit of 5 and found that matches lasted an average of 2:30 min. As a result, we decided to increase the FragLimit in order to achieve an average match duration of closer to 5 min.

5.1.5 Measures

We were able to automatically record match information in a database using our framework, allowing us to quickly handle it using queries. The map utilised, the duration of the match, the match winner, the score of the two players, as well as the number of times they fragged, committed suicide, or killed their opponent are all collected for each game. We gathered the participants' assessments as well as their degree of conviction after each match, which allowed us to create two scores: a humanness score and a reliability score. The score increases if the player is found to be human; otherwise, it lowers. If the given degree of confidence was 0, the score stayed unchanged (i.e., "not sure at all"). Only human players have a reliability score since machines do not judge. This score is enhanced when the player correctly assesses his opponent; otherwise, it is lowered. At the completion of the

[2]http://afia-competitions.fr/botcontest/.

[3]A frag is a video game term equivalent to "kill", with the main difference being that the player can re-spawn (reappear and play again).

[4]Translated from French.

TABLE 2 | Competition results.

Teams	Humanness	Humanness with certainty
Humans' avg.	0.38	3.08
A Human Guy	−0.19	−1.67
Communaute de Nao	−0.29	−2.59
AOP	−0.33	−3.25

study, participants were requested to complete a four-question questionnaire to measure their video game expertise:

1) How often do you play video games?

Everyday, Several times a week, Only on weekends, A few times a month, Only during holidays, Never.

2) What device do you use to play video games?

Computer, Console, Hand-held game console, Arcade game, Other device.

3) What types of games do you play?

First-Person Shooter, Strategy games, Platform games, Adventure, Action Games, Role Playing Game, Educational games, Management Games, Simulation games, Sports Games, Racing Games, MMORPG, Massively multi-player on-line role-playing game, Physical or sports games.

4) Do you play:

Alone, With computer players (or bots), On-line with strangers, On-line with friends or family, With physically present players.

For question 1., participants could only choose one answer and for questions 2. and 4., they could select multiple answers. For question 3., they had to select only the type of games they play and sort them from most to less often.

5.2 Results

The competition results are listed in **Table 2**. We should notice that the bots' scores are all negative, indicating that none of them passed the exam. Whether we use the confidence scale or not, the resulting ranking is the same. We decided to present only the results using the humanness score without the degree of certainty in this section for the sake of simplification because we used this scale for the simple purpose of discouraging participants from not doing the cognitive work (as explained in **section 4**) and because analysing data with and without the degree of certainty gave us the same results, we decided to present only the results using the humanness score without the degree of certainty in this section for the sake of simplification. We used a *t*-Test to examine the difference in humanness scores between people and bots, which yielded a $p < 0.001$, suggesting that the difference is significant.

We were able to investigate several characteristics of the procedure using the data we acquired throughout the competition. First and foremost, the bar plot in **Figure 3** depicts the match duration distribution for each map. Matches were divided into five categories: matches lasting less than 2 minutes, 3 minutes, 4 minutes, and 5 minutes; and games concluding with a time-limit condition of 5 minutes. It's worth noting that the length of the match varies from one map to the next. A Kruskal-Wallis test.[5] was used to confirm this result, with a $p < 0.001$ suggesting that the mean match duration changes considerably depending on the maps. This backs up our conclusions from the pre-tests: on certain maps, players see their opponents far more frequently than on others. The humanness score changes depending on the map, although bots score higher than humans (see **Figure 4**). Bots had $p = 0.093$ while humans had $p = 0.52$ in the Kruskal-Wallis test. As a result, the bots' humanness score changes greatly depending on the map. However, because the length of the match is determined by the map, we must proceed with caution when interpreting these outcomes. According to the bar plot in **Figure 5**, the humanness score appears to fluctuate with the duration of the matches: the shorter the matches, the lower the score. The Kruskal-Wallis test, however, yielded $p = 0.39$ for bots and $p = 0.38$ for humans, preventing us from rejecting the null hypothesis. We also looked at a possible link between the player's humanness score and 1) whether or not he won, 2) his score, and 3) how many times he died as a result of his own acts. Kruskal-Wallis test were 1) $p = 0.67$, 2) $p = 0.52$, and 3) $p = 0.76$. This prevents us from rejecting the null hypothesis, leading us to believe that there is no relationship between these factors and the humanness score.

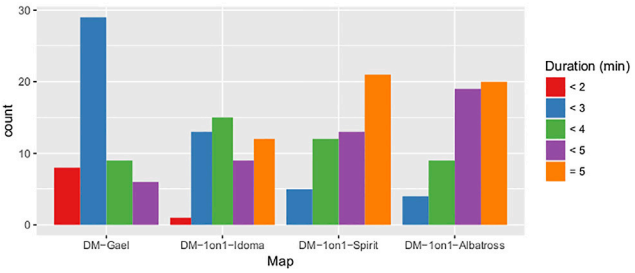

FIGURE 3 | Bar plot of the match duration (in minutes) depending on the maps.

We looked at four aspects of gaming behaviours using the final questionnaire: 1) the regularity with which people play games, 2) the kind of games they typically play, 3) the devices they usually use, and 4) the types of gamers they usually encounter. We divided the participants into three levels of competence based on their dependability ratings. Because several of the participants had the same intermediate score, we divided them into the following categories: 10 best–40 intermediate–10 worst. The

[5]This test is a non parametric alternative to the One-Way ANOVA and is used when the dependant variable does not meet the normality assumption. It can be used to assess for significant differences on a dependent variable by a categorical independent variable (with two or more groups).

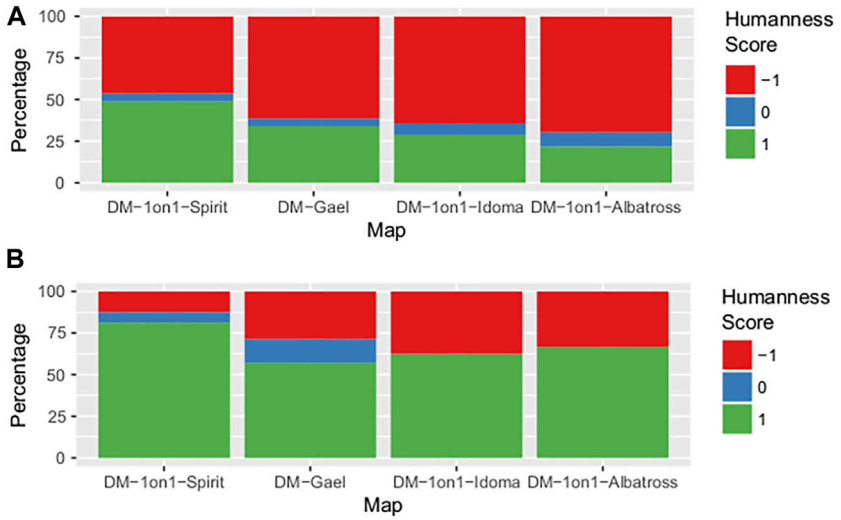

FIGURE 4 | Bar plot of the humanness score for **(A)** bots and **(B)** humans depending on the maps.

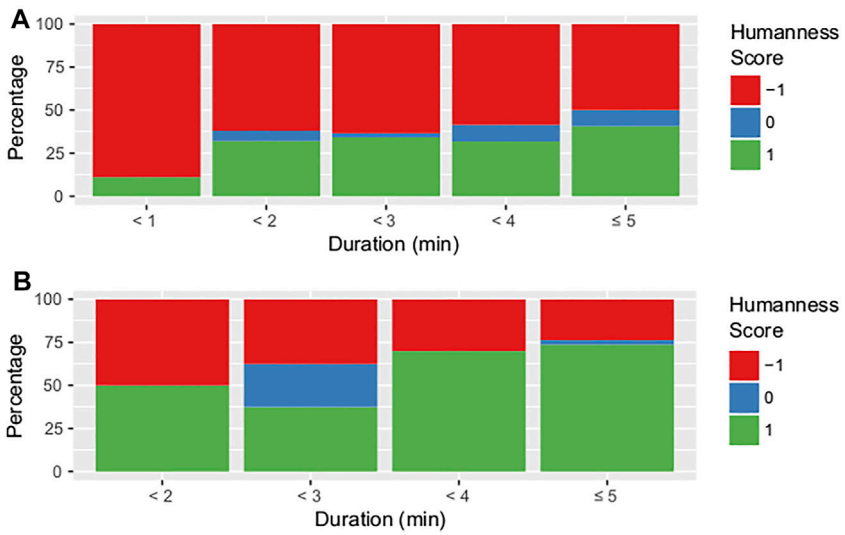

FIGURE 5 | Bar plot of the humanness score for **(A)** bots and **(B)** humans depending on the match duration (in minutes).

top judges correctly recognised all of their opponents, whereas the worst judges were wrong at least three times out of four.

We created a contingency table to see if there was a link between the participants' degree of competence and their gaming frequency. We cannot reject the null hypothesis since the chi square of independence between the two variables is equivalent to 11.74 (p = 0.3). The correspondence analysis, on the other hand, yielded a surprising conclusion (see **Figure 6**) is rather interesting since it reveals that the top judges play every day, the worst judges never play, and intermediate judges only play once in a while.

We achieved a chi square of independence between the two variables of 31.60 (p = 0.024) using the same approach as before. As a result, we can rule out the null hypotheses and conclude that the degree of skill of the participants is related to the sort of video game they often play. The correspondence analysis' outcome (see **Figure 7**) enables us to get additional information about this reliance The red letters on the diagram relate to the types of games listed in **section 5.1.5**. Participants with the greatest degree of experience play games like (A) first-person shooter games and (D) adventure and action games, as seen in this graph. Shooting and combat are key elements in both of these games. Intermediate-level judges play games such as (B) strategy games, (E) role-playing games, and (C) platform games. Combat stages are fairly prevalent in these sorts of games, although they are not a major component of the game. Games

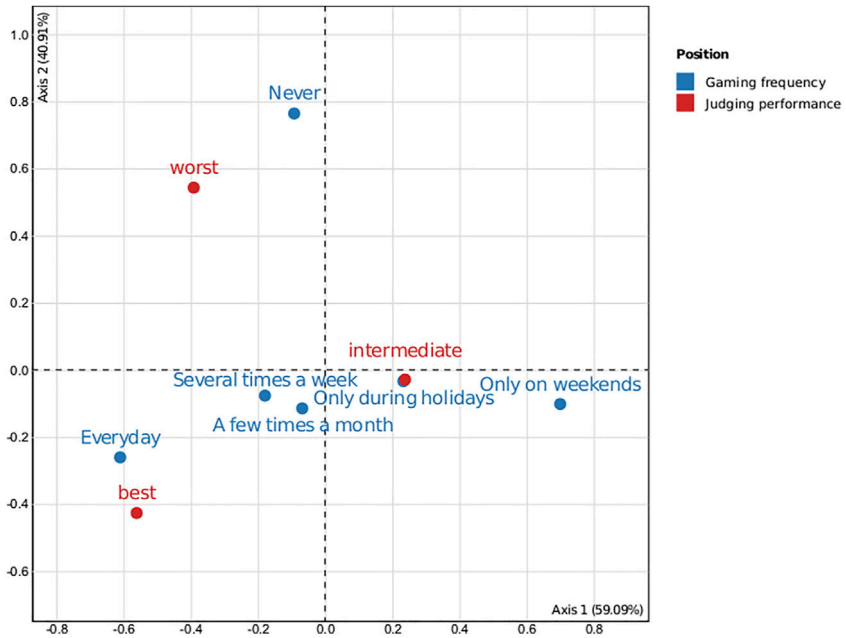

FIGURE 6 | Correspondence analysis factor map.

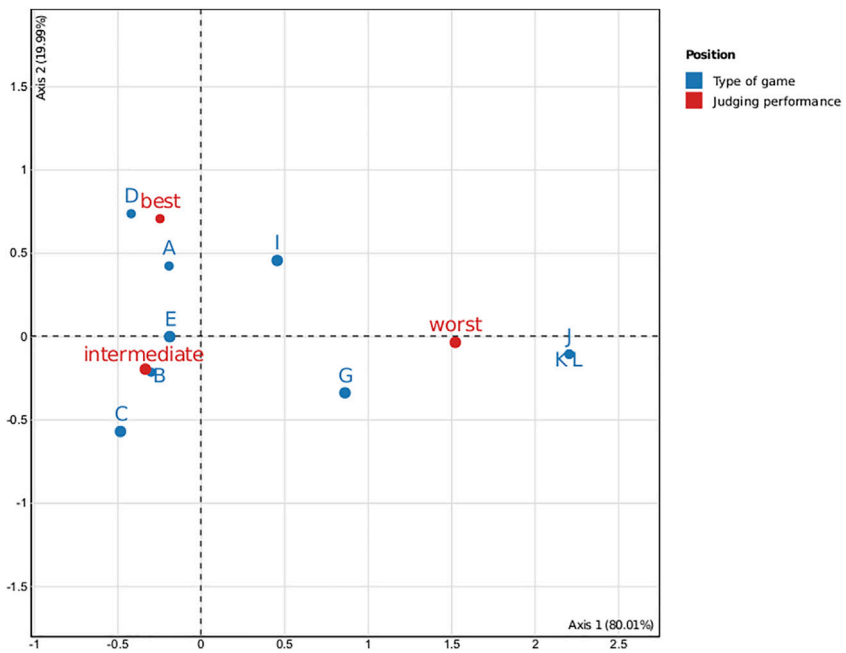

FIGURE 7 | Correspondence analysis factor map.

like (I) Sports Games, (J) Racing Games, (K) MMORPG, and (G) Management Games are preferred by participants with the lowest degree of competence. Shooting stages are not common in these sorts of games, or they are quite infrequent.

For all levels of competence, the distribution of the responses chosen by the participants about the equipment utilised is comparable (see **Table 3**). There is therefore no link between these two elements.

The response distribution for each level of skill is shown in **Table 4**. We've seen that the participants with the highest degree of competence are the ones who, unlike the others, prefer to face a variety of players. We used a multiple correspondence analysis to

TABLE 3 | Distribution of the devices usually used to play according to the level of expertise (in percentage).

Judging level	Computer	Console	Handheld	Arcade	Phone
Best	90	50	40	0	40
Intermediate	80	38	10	7	33
Worst	60	40	10	0	40

TABLE 4 | Distribution of the type of players usually met in games according to the level of expertise (in percentage).

Judging level	Alone	Bots	Strangers	Friends	PPP[6]
Best	100	60	70	70	90
Intermediate	80	33	28	48	48
Worst	70	40	40	70	40

back up our observations. In a Euclidean space, this technique locates all categories. The first two dimensions of Euclidean space (see **Figure 8**) are displayed to evaluate the connections between the categories. On this graph, 1 denotes a favourable response (i.e., the participant claimed to have previously played with this sort of player), while 0 denotes a negative response. The positive values are on the left side of the graph, while the negative values are on the right. The best judges are on the left side of the graph, while the poorest and intermediate judges are on the right side. This demonstrates that the values on the right are more widely shared among the participants with the highest degree of knowledge than among the rest, confirming our findings from **Table 4**.

5.3 Discussion

This research allowed us to make some interesting findings about the competition's features as well as the competitors' levels of skill. To begin, we discovered that the number of times the players encounter is determined by the map utilised in the contest. Furthermore, bots are seen as more human-like on some maps than others, thus varied behaviour may be predicted depending on the environment. The battles on the DM-Gael map, for example, are fast-paced, which is understandable given that it is made up of a single room where hiding is extremely tough. As a result, close fighting is more frequent than sniping on this sort of battlefield. In order to notice these varied methods, it appears that integrating different maps while judging the believability of the bots is crucial. We also noted that the player's humanness score is unaffected by his score or whether he has won or lost. This is especially intriguing because player performance and believability appear to be unrelated. The results of the experiment allowed us to profile the participants who had the best level of expertise in distinguishing bots from human players: players who primarily play games with shooting or fighting as a main component, and players who are used to playing against a variety of opponents, including bots, strangers, and physically present players (they also tend to play games regularly). Participants with the lowest degree of competence are more likely to play games without any fighting, either alone or with friends or family. These players do not have a sufficient understanding of the sort of game utilised in the tournament to adequately assess their opponents. Even while the game's rules are straightforward (kill your opponent as many times as possible), mastering this sort of game requires extensive training. Despite the addition of a training phase, we found that some participants who had never played a game like this previously struggled to navigate the environment. Certain behaviours, such as opponents leaping after being noticed even when there were no barriers, startled several of these gamers. However, because it is more difficult to hit a leaping adversary with a headshot in a first-person shooter, this behaviour is common. Players will expect different behaviour based on their level of knowledge, demonstrating the subjectivity of believability.

[6]Physically Present Players.

6 EXPERIMENT 3: REPORTING SUSPECTED CHEATERS

Until now, one of the main problem is that the gameplay can be modified by these so-called "first-person" assessment methods, as we have seen in the previous sections. Players are more focused on judging than playing the game, which introduces new behaviours in the game. In this section, we propose a new method to indirectly assess bots' believability with both an objective and subjective evaluation. With this approach, the gameplay is not affected since the game is played normally and players are not asked to judge their opponents.

While some constructs (i.e., the characteristic to assess, so in our case: the believability of a bot) can be measured directly, others require more subtle or indirect measurement. Prior research provides a valuable context for work on measuring a construct (Cronbach and Meehl, 1955; Campbell and Fiske, 1959). Current methods of assessing constructs can be informed by drawing on the successes of prior efforts. However, if they have consistently failed to yield expected results, it may indicate the need to strike off on a different path in order to evaluate the construct. This is the solution we have adopted and we have sought to put in place a protocol for assessing the believability through indirect measurements.

To do this, we were inspired by the reporting systems present in most online multiplayer video games. These systems are used by players to report prejudicial behaviours faced when playing a game. Most of the time, these systems offer many options to report abuses, but these options may differ depending on the type of game and the device used to play. On home consoles for instance, it can be difficult, if not impossible, to install third party software that would allow a player to cheat while this manoeuvre is rather simple on a computer. Therefore it is more likely to find an option to report cheating on PC games rather than home consoles games. The options that are generally present in any games and devices are: harassment, offending language or name and being "away from the keybord". In certain games where the collaboration between the members of a team is essential, one can find reporting reasons such as "poor team work" or "team

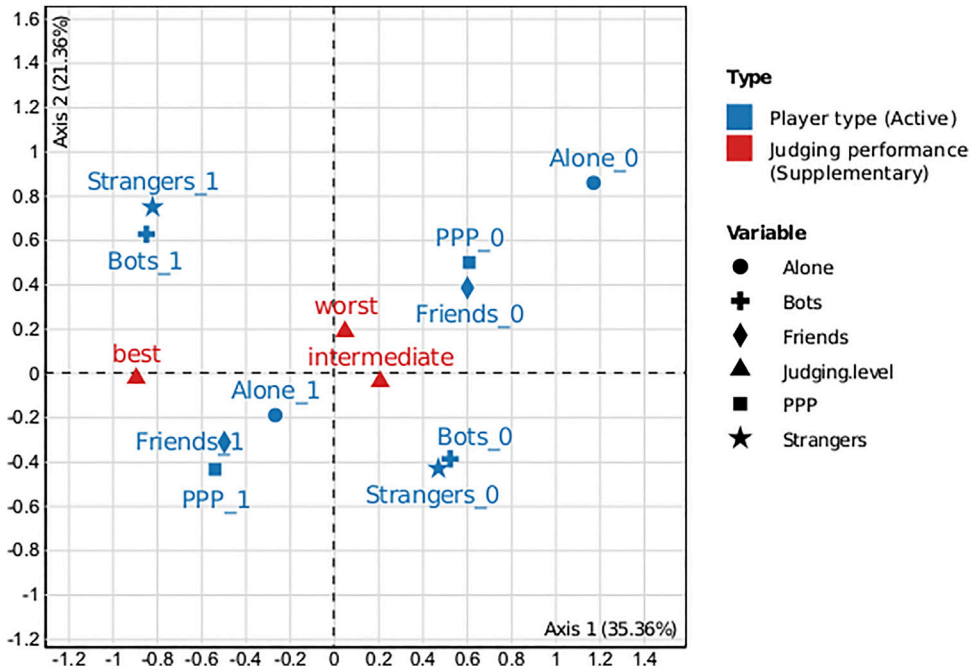

FIGURE 8 | Multiple correspondence analysis plot for dimensions 1 and 2.

damage" for instance. Once the game company has been warned of the harmful behaviour, it can decide the penalty to give to the player. This can range from a simple warning to several days of no play or even to the total closure of the account.

Our proposal consist in adding options to the reporting form to allow players to signal the presence of bots. We hypothesised that the more often a bot is reported, the less believable it is. Indeed, we assumed that the bot that will be most reported will be the one whose behaviour is the most different from the expected behaviour in the game and therefore the least believable. This allows us to evaluate the believability of the bots objectively (Yan and Randell, 2005; Alayed et al., 2013).

6.1 Method

To validate our approach we conducted an experiment where we invite participants to fill a questionnaire after playing a succession of matches with our reporting system. We wanted to verify if the bot that was reported the most often was the one that is deemed the least believable by the participants.

6.1.1 Participants

Ads were placed in different parts of the city to recruit the participants. They were all volunteers and no compensation was provided for their participation. Seventeen participants including sixteen men (94.1%) and one woman (5.9%) took part in the experiment. For all the participants, French was their native language. Their mean age was 28, ranging from 19 to 42 years old. 47.1% of participants reported playing every day, 17.6% play several times a week and 11.8% play a few times a month. Among the participants, 17.7% consider themselves as novice players, 58.8% as amateurs and 23.5% as experts. All data were analysed anonymously and all participants gave written informed consent prior to participation.

6.1.2 Material

Since we already had the bots and the system to manage clients and servers automatically for the video game UT2004, we chose to use it again for this last experiment. However, this game does not include a reporting system by default. We therefore developed a reporting system for this game by taking inspiration from existing ones in other video games. Various solutions exist to access the reporting form. The three most common solutions are:

- Right click on the player's avatar in the game window.
- By right clicking on the name of the player in the chat.
- In the game menu by choosing the player from a list.

Since we have disabled the chat as we are not trying to evaluate the bot's ability to communicate, we can not use the second option. The first solution is not suitable for a game such as a FPS. Indeed, this type of game having a very fast pace, it is difficult for the player to perform a manipulation in the game without becoming an easy target. Therefore, the third solution seemed to us to be the most suitable. To facilitate the use of the reporting form, we integrated it into a web page that can be positioned next to the game window. To access it, the player simply has to change the active window with the keys combination.

To determine options we would integrate into the reporting form, we studied the codes of conduct of several video games (Call of Duty Black Ops 3, Call of Duty World War II and Tom Clancy's Rainbow 6 Siege, ...), existing reporting forms (Overwatch, League of Legends and World

of Warcaft, . . .), as well as previous studies (Yan and Randell, 2005; Alayed et al., 2013). The ten most popular reasons to report a player are:

1) Spam
2) Bug exploitation
3) Automatic aiming and shooting
4) Alteration of wall texture
5) Using bots
6) Aggressive language
7) Inappropriate name or profile picture
8) Personal statistics modification
9) Fraud
10) Harassment

Since the chat is disabled, all options related to this activity have been removed (i.e., 1, 6, 7, 10). The game does not include items that can be purchased with real money so we also removed the option for fraud (9). In order to adapt the eighth option to the game in question, we divided it into two sub-categories. The first can be used to report the increase of the resistance to damage inflicted by other players while the second reports the increase of the damage caused by the weapon of the cheater. If no option is appropriate for the player, he/she is free to choose the "Other" option and fill the field with the desired reason. Here is the list of options we chose to use for the experimental reporting form:

1) Bug exploitation
2) Automatic aiming and shooting
3) Alteration of wall texture
4) Using bots
5) Increase of damage resistance
6) Increase of weapon damage
7) Other

The goal of this new approach was to stay as close as possible to the way the game is normally played. Generally, people wishing to play UT2004 would connect to a server, and start to play a succession of matches once a minimum number of players have logged in. They play against several players at once and meet on several maps of the game. It was important for us to replicate this experience. Fortunately, because of the flexibility of our computer system, it was particularly simple to put this in place. The game engine already has a system to change maps automatically at the end of each game by default. We used our system to start the servers and connect the players. The game engine then took care of starting the game matches successively as it normally would.

6.1.3 Procedure

The experiment had two conditions: a control condition and an experimental condition. In the control condition, the four participants played all against each other without any bots. In the experimental condition, the four participants were divided into two groups. Each participant would play against the other member of the group and two bots. The two bots were the ones who came first (*A Human Guy*) and third (*AOP*) in the BotContest competition.

Participants were welcomed and invited to take place at one of the computer dedicated to the experiment. The same physical arrangement was used as in the two previous experiments. Participants were only informed that it was an experiment on the reporting forms in video games and that some participants might have access to a cheat technique during the game. In fact none of them had access to such a feature. It was just a pretext to instigate them to use the report form.

After filling and signing a consent form, participants were directed to the questionnaire used to evaluate their gaming habits. Then, as with previous models, participants started with the tutorial (which we have not changed). Then, the participants could start playing the game. They had to play four matches of 5 minutes each. The instruction was to arrive at the maximum score as quickly as possible while using the reporting form when observing suspicious behaviours. We set the maximum score to 30 because this score is difficult to reach within the time limit but not impossible. Thus participants must fully invest them-self into the match to have a chance to reach this score. The matches followed one another automatically and a different map was used for each of them. Once the game session was over, the participants had to fill a final questionnaire. This questionnaire made it possible to collect information on the participants experience with the report form as well as their opponents. The first part of this questionnaire only served as a distraction and allowed not to focus only on the opponent. The second part of the questionnaire allows us to collect data on the gaming experience and the perception, or not, of the presence of the bots by the participants.

6.1.4 Variables and Measures

Number of reports and the reason of reporting were recorded. Participants had to judge the believability of bots with a 6 points Likert scale, going from 1 "not believable at all" to 6 "very believable". Participants were asked to indicate how many human players and bots they thought they faced. They could choose a value between 0 and 3. Participants were also asked to specify their degree of certainty regarding the previous answer (number of human players and bots). They could choose their answer on a Likert scale (going from 0 "not sure at all" to 6 "completely sure").

6.2 Results

The participants in the control group used the reporting form on average 1 time, while those in the experimental group reported on average 2.7 times (see **Figure 9A**). The bivariate Wilcoxon test gave a $p = 0.055$ which does not allow us to reject the null hypotheses. However, we can see that this p-value is very close to being significant. We can therefore conclude that a difference between the two groups seems to be emerging and that the experimental group tends to signal more often than the control group.

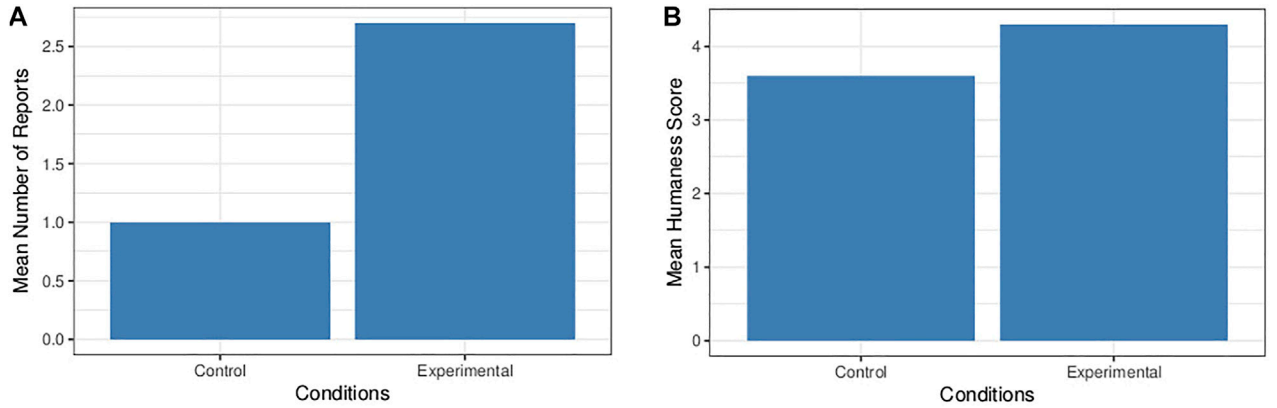

FIGURE 9 | Mean number of reports depending on the condition **(A)**. Mean humanness score for bots depending on the condition **(B)**.

In the experimental condition, 85.2% of the reports were for a bot, out of which 68.2% were for *A Human Guy* and 34.8% for *AOP*. We analysed the possibility of a difference in the number of reports between the two bots. *A Human Guy* (1.4 ± 1.17) has been reported twice as often as *AOP* (0.7 ± 0.67), however, the difference between the two is not significant according to a Wilcoxon test (V = 17, $p = 0.202$).

The different reasons of reporting have been studied to see if some of them were chosen more often. The Fisher exact Test seems to reveal that some were used more than others ($p = 0.034$). The reasons "Increase of damage resistance", "Automatic aiming and shooting" and "Using bots" seem to be chosen more frequently than then other ones and the "Other" option was never used.

The experimental group found that bots were rather believable (4.3 ± 1.4). The same question was asked to the control group, even though there were no bots present in this condition. They thought that bots were believable on average (3.6 ± 1.4). In **Figure 9B**, the two groups do not seem to be significantly different, this was confirmed by a Wilcoxon test which gave a $p = 0.311$.

Regarding how many human players participants thought they faced answers do not seem to be significantly different ($p = 0.954$) between the control group (2.57 ± 0.53) and the experimental group (2.6 ± 0.52). Regarding the number of bots, the difference is not significant ($p = 0.869$) between the control group (0.71 ± 0.76) and the experimental group (0.6 ± 0.62).

Participants' degree of certainty for the number of human player do not seem to be significantly different ($p = 0.547$) between the control group (2.71 ± 1.98) and the experimental group (3.4 ± 1.95). The same question was asked regarding the number of bots. Again, the difference is not significant ($p = 1$) between the control group (2.57 ± 1.9) and the experimental group (2.7 ± 2.16).

A Pearson Correlation test was performed to study an eventual link between the number of reports and the believability score for bots. The control group shows signs of a negative correlation ($p = 0.058$, cor = −0.739 1) whereas for the experimental group (see **Figure 10**), a strong negative correlation seems to appear between the number of reports and the believability score ($p = 0.027$, cor = −0.689 8).

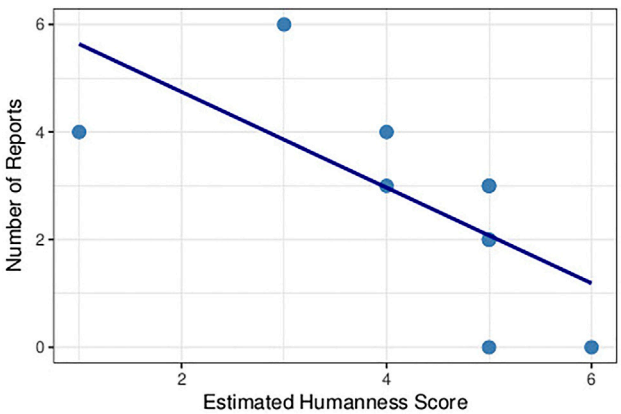

FIGURE 10 | Negative correlation between the bots' estimated humanness and the number of reports in the control group.

We also studied the usability of our reporting form. Regarding the complexity of manipulation to perform to access the form: 5.9% of participants found it complex, 17.6% found it quite simple, 17.6% found it simple and 58.8% found it very simple, which is very satisfying. In addition, 82.3% of participants reported being ready to use this type of reporting form if they had the opportunity.

6.3 Discussion

Despite the fact that our population is relatively small (10 participants for the experimental condition and 7 for the control one), our statistical analysis gave very encouraging results.

Firstly, we can see that a significant difference seems to appear between the experimental group and the control group with regard to the number of reports made. Participants in the experimental group would tend to report more often than those in the control group (almost three times more often on average). Furthermore, we can see that in the experimental

condition, bots are reported five times more often than human players. This could reflect a difference in behaviour between human players and bots. We have deliberately incorporated different reasons into the reporting form which could lead to improvements for the implementation of the bots. For example, the bot *A Human Guy* was reported four times for "Automatic aiming and shooting" and three times for "Alteration of wall texture", "Increase of damage resistance" and "Using bots". The first two reasons suggest that the bot's firing behaviour could be improved. The other two, on the other hand, give less indications for improvements. The third reason might suggest that the bot is too efficient at collecting health points which could give him the illusion of having more resistance.

The second element of our statistical study which is particularly interesting is the measurement taking into account both objective data (number of reports) and subjective data (humanness score). This has never been used together before for assessing the believability of bots and that is the particularity of our approach. The statistical analysis seems to reveal a negative correlation between those two variables, and particularly in the experimental condition where the correlation is strong. This result is particularly encouraging since it seems to show that our goal is achieved. Indeed, we have been able to set up an evaluation of the believability of the bots which allows to play the game as it should be without having an impact on the gameplay and which makes it possible to obtain an indication on the believability of the bots as well as suggestions for improvement.

However, this study has some limitations, such as the number of participants ($n = 17$), which limits the interpretation of the statistics performed. Parametric tests, such as the Student's t-test, are more powerful than non-parametric tests, i.e., the probability of rejecting the null hypotheses is higher. However, certain criteria must be respected in order to carry out these parametric tests (Elliott and Woodward, 2007; Cronk, 2017), such as having a normal distribution, or having equal variances for the two populations. It is therefore preferable to have a large population size ($n \geq 30$) in order to increase the possibility of a normal distribution of the data and an homogeneity of the variances (Ghasemi and Zahediasl, 2012). It would be interesting for future experiments to have more participants in order to be able to perform parametric tests and thereby deepen, and perhaps strengthen, the results obtained during this experiment.

We found that it would be possible to slightly improve the last questionnaire of the experiment so as to evaluate the bots' believability individually. During this experiment, participants were not asked to evaluate each of their opponents' believability but rather, they were asked to mention the number of bots they thought they faced, their degree of certainty, and whether the bots they faced seemed believable. There is therefore no real distinction between the individual players during the evaluation. A distinction could have helped us to conduct further analysis and investigate the existence of a direct link between the number of reports and the humanness score for each bot.

The results we obtained in this study do not match the ranking of the BotContest competition presented in the previous section. Indeed the bot *A Human Guy*, winner of the competition, was reported more often than the bot *AOP*. This reverse ranking did not surprise us. Indeed, the bot *A Human Guy* being based on a mirror mechanism, is perfect for a situation where the gameplay is changed by the judgement. Because the bot imitates the judges, they may be led to think that the player in front of them is also judging or trying to communicate. The bot *AOP* however has been developed to play the game as it is supposed to be played. It seems normal to us that the bot *A Human Guy* was judged more believable in the context of the competition where the judgement of the believability was an important element of the gameplay.

7 CONCLUSION AND FUTURE WORK

The goal of this article was to put in place a rigorous protocol to evaluate the believability of computer players in multiplayer video games. This notion of believability is particularly complex to evaluate due to its subjectivity. Indeed, gamers will not perceive believability in the same way according to their familiarity with the video game and their level of expertise in it. To propose a new protocol, we embarked on a system of trial and error, each new protocol drawing on the successes of its predecessor whilst eliminating the failures.

Firstly, we conducted a literature review of the protocols previously used to assess the believability of computer players. After analysing them in detail, we identified seven features that characterise the assessments and which vary significantly from one to another. We discussed that when designing a new protocol, these features need to be chosen carefully in order to not introduce a bias into the evaluation. After an in-depth analysis of these protocols, we gave recommendations for the features that are well established. We also identified the other features that still need further study and testing to be determined. During the literature review we found out that the video game's gameplay could be affected by the assessment process. To avoid this we sought to hide the purpose of the evaluation by building a questionnaire aiming attention at several aspects of the game. The goal being to disperse the attention of the participants on the whole game rather than simply on their opponent. Throughout our study we used the video game Unreal Tournament 2004, a first person shooter game, since it has been used many times in previous studies (Bída et al., 2012). To facilitate the execution of the evaluation, we developed a system that partially automates the evaluation process. It is responsible for running the game servers and for automatically connecting players and bots to it. This system proved to be effective and flexible since it has also been used successfully for the implementation of the two other protocols that we proposed. Our first protocol having given unconvincing results, we wondered if this could be due to the level of expertise of participants in video games. We tried out our protocol during a conference, during which we organised a competition. We took advantage of this event to profile the judges according to their ability to correctly distinguish bots from human players. We found that the best judges are players who mainly play games that have shooting or fighting as their

main component and players who are used to playing against different types of opponents including, in particular, bots, strangers and physically present players (they also tend to play games regularly). On the other hand, the judges with the lowest level of expertise tend to play games that do not include combat at all and usually play alone or with friends or family. These observations showed us that the level of the players can have an influence on their expectations concerning the behaviours of their opponents. It therefore seems important to integrate players of different levels in the evaluation in order to obtain consistent results. Finally, from the observations that we could make during our previous experiments, we came up with a completely new design. For this new approach we tried to use the game as it is normally played, with the aim of minimising as much as possible the impact of the assessment on the gameplay. We decided to take inspiration from the reporting systems already present in many video games. We propose to create a reporting form that includes options for reporting undesirable behaviours that may be manifested by bots. Our proposal is therefore to evaluate the believability of bots indirectly by using an objective measure: the number of reports made against the bot. We conducted an experiment to validate our approach and obtained promising results. In particular, our statistical analysis showed that there is a negative correlation between the number of reports and the believability of the bots, which meets our hypothesis.

Our new protocol makes it possible to evaluate the believability of the bots while respecting the gameplay of the game and by involving players with different levels of expertise, which is a hefty improvement compared to the previous evaluation methods. However, many improvements are still possible. Our protocol can easily adapt to different video game genres such as, action, strategy, role-playing or sports games. However, for this, different reporting options should be proposed depending on the game genre. One way to improve our protocol would be to study the harmful behaviours, and more particularly those associated with bots in video games of different genres. This would help to establish lists of reporting options for each game genre, which would make it easier to set up an evaluation for any video game that is not a first person shooter.

DATA AVAILABILITY STATEMENT

The UtBotEval software is freely available at https://git.enib.fr/even/utboteval and https://git.enib.fr/even/ut2004-utboteval-mod.

ETHICS STATEMENT

The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CE, AB, and CB contributed to conception and design of the study. CE developed the software and performed the statistical analyses. CE performed the user evaluation. AB and CB were the project administrators and supervised the work. All authors contributed to article revision, read, and approved the submitted version.

REFERENCES

Acampora, G., Loia, V., and Vitiello, A. (2012). Improving Game Bot Behaviours through Timed Emotional Intelligence. *Knowledge-Based Syst.* 34, 97–113. doi:10.1016/j.knosys.2012.04.012

Alayed, H., Frangoudes, F., and Neuman, C. (2013). "Behavioral-based Cheating Detection in Online First Person Shooters Using Machine Learning Techniques," in Computational Intelligence in Games (CIG), 2013 IEEE Conference on (Citeseer) (IEEE), 1–8. doi:10.1109/cig.2013.6633617

Arrabales, R., Ledezma, A., and Sanchis, A. (2012). *ConsScale FPS: Cognitive Integration for Improved Believability in Computer Game Bots*. Berlin, Heidelberg: Springer Berlin Heidelberg, 193–214. doi:10.1007/978-3-642-32323-2_8

Arrabales, R., Ledezma, A., and Sanchis, A. (2010). ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents. *J. Conscious. Stud.* 17, 131–164.

Bailenson, J. N., Aharoni, E., Beall, A. C., Guadagno, R. E., Dimov, A., and Blascovich, J. (2004). "Comparing Behavioral and Self-Report Measures of Embodied Agents' Social Presence in Immersive Virtual Environments," in Proceedings of the 7th Annual International Workshop on PRESENCE (IEEE), 216–223.

Bates, J. (1994). The Role of Emotion in Believable Agents. *Commun. ACM* 37, 122–125. doi:10.1145/176789.176803

Bevacqua, E., Stanković, I., Maatallaoui, A., Nédélec, A., and De Loor, P. (2014). "Effects of Coupling in Human-Virtual Agent Body Interaction," in *Intelligent Virtual Agents* (Springer), 54–63. doi:10.1007/978-3-319-09767-1_7

Bída, M., ČernŸ, M., Gemrot, J., and Brom, C. (2012). "Evolution of Gamebots Project," in International Conference on Entertainment Computing (Springer), 397–400.

Bogdanovych, A., Trescak, T., and Simoff, S. (2016). What Makes Virtual Agents Believable? *Connect. Sci.* 28, 83–108. doi:10.1080/09540091.2015.1130021

Bossard, C., Benard, R., De Loor, P., Kermarrec, G., and Tisseau, J. (2009). "An Exploratory Evaluation of Virtual Football Player's Believability," in Proceedings of 11th Virtual Reality International Conference (VRIC'09) (IEEE), 171–172.

Bosse, T., and Zwanenburg, E. (2009). "There's Always hope: Enhancing Agent Believability through Expectation-Based Emotions," in 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (IEEE), 1–8.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychol. Bull.* 56, 81–105. doi:10.1037/h0046016

Coleridge, S. T. (1817). *Biographiae Litteraria of Biographical Sketches of My Literary Life and Opinions*. London: Rest Fenner, 23 Patermaster Row.

Cronbach, L. J., and Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychol. Bull.* 52, 281–302. doi:10.1037/h0040957

Cronk, B. C. (2017). *How to Use SPSS®: A Step-by-step Guide to Analysis and Interpretation*. London: Routledge.

Dolnicar, S., Grün, B., and Leisch, F. (2011). Quick, Simple and Reliable: Forced Binary Survey Questions. *Int. J. Market Res.* 53, 231–252. doi:10.2501/ijmr-53-2-231-252

Elliott, A. C., and Woodward, W. A. (2007). *Statistical Analysis Quick Reference Guidebook: With SPSS Examples*. London: Sage.

Even, C., Bosser, A.-G., and Buche, C. (2017). "Analysis of the Protocols Used to Assess Virtual Players in Multi-Player Computer Games," in 14th International Work-Conference on Artificial Neural Networks (IEEE), 657–668. doi:10.1007/978-3-319-59147-6_56

Even, C., Bosser, A.-G., and Buche, C. (2018). "Bot Believability Assessment : a Novel Protocol & Analysis of Judge Expertise," in 17th International Conference on Cyberworlds (CW) (IEEE), 96–101. doi:10.1109/cw.2018.00027

Ghasemi, A., and Zahediasl, S. (2012). Normality Tests for Statistical Analysis: a Guide for Non-statisticians. *Int. J. Endocrinol. Metab.* 10, 486–489. doi:10.5812/ijem.3505

Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Goffman, E. (1963). *Behavior in Public Places: Notes on the Social Organization of Gatherings*. New York: Free Press.

Gorman, B., Thurau, C., Bauckhage, C., and Humphrys, M. (2006). Believability Testing and Bayesian Imitation in Interactive Computer Games. *From Anim. Animats* 9 (1), 655–666. doi:10.1007/11840541_54

Heeter, C. (1992). Being There: The Subjective Experience of Presence. *Presence: Teleoperators & Virtual Environments* 1, 262–271. doi:10.1162/pres.1992.1.2.262

Hingston, P. (2010). "A New Design for a Turing Test for Bots," in Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games (IEEE), 345–350. doi:10.1109/itw.2010.5593336

Hingston, P. (2009). A Turing Test for Computer Game Bots. *IEEE Trans. Comput. Intell. AI Games* 1, 169–186. doi:10.1109/tciaig.2009.2032534

Hinkkanen, T., Kurhila, J., and Pasanen, T. A. (2008). "Framework for Evaluating Believability of Non-player Characters in Games," in *AI and Machine Consciousness* (Springer).

Koehler, D. J., and Harvey, N. E. (2004). *Blackwell Handbook of Judgment and Decision Making*. Malden, MA, United States: Blackwell Publishing.

Krosnick, J. A. (2002). The Causes of No-Opinion Responses to Attitude Measures in Surveys: They Are Rarely what They Appear to Be. *Surv. Nonresponse* 1, 87–100.

Laird, J. E., and Duchi, J. C. (2001). "Creating Human-like Synthetic Characters with Multiple Skill Levels: A Case Study Using the Soar Quakebot," in Papers from 2001 AAAI Spring Symposium, Artificial Intelligence and Interactive Entertainment I (Springer), 54–58.

Le Hy, R., Arrigoni, A., Bessière, P., and Lebeltel, O. (2004). Teaching Bayesian Behaviours to Video Game Characters. *Robotics Autonomous Syst.* 47, 177–185. doi:10.1016/j.robot.2004.03.012

Livingstone, D. (2006). Turing's Test and Believable AI in Games. *Comput. Entertain.* 4, 6. doi:10.1145/1111293.1111303

Llargues Asensio, J. M., Peralta, J., Arrabales, R., Bedia, M. G., Cortez, P., and Peña, A. L. (2014). Artificial Intelligence Approaches for the Generation and Assessment of Believable Human-like Behaviour in Virtual Characters. *Expert Syst. Appl.* 41, 7281–7290. doi:10.1016/j.eswa.2014.05.004

Loyall, A. B. (1997). *Believable Agents: Building Interactive Personalities*. Carnegie Mellon University. Ph.D. thesis.

Lucas, S. M., Mateas, M., Preuss, M., Spronck, P., and Togelius, J. (2012). Artificial and Computational Intelligence in Games (Dagstuhl Seminar 12191). *Dagstuhl Rep.* 2, 43–70.

Mac Namee, B. (2004). *Proactive Persistent Agents: Using Situational Intelligence to Create Support Characters in Character-Centric Computer Games*. Trinity College: University of Dublin. Ph.D. thesis.

MacLean, C. L., and Dror, I. E. (2016). "A Primer on the Psychology of Cognitive Bias," in *Blinding as a Solution to Bias* (Academic Press), 13–24. doi:10.1016/b978-0-12-802460-7.00001-2

Magnenat-Thalmann, N., Kim, H., Egges, A., and Garchery, S. (2005). "Believability and Interaction in Virtual Worlds," in Proceedings of the 11th International Multimedia Modelling Conference (IEEE), 2–9.

McGlinchey, S., and Livingstone, D. (2004). "What Believability Testing Can Tell Us," in Proceedings of the International Conference on Computer Games: Artificial Intelligence, Design, and Education (IEEE), 273–277.

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., and De Carolis, B. (2005). "Greta. A Believable Embodied Conversational Agent," in *Multimodal Intelligent Information Presentation* (Springer), 3–25. doi:10.1007/1-4020-3051-7_1

Polceanu, M. (2013). "Mirrorbot: Using Human-Inspired Mirroring Behavior to Pass a Turing Test," in IEEE Conference on Computational Intelligence in Games (CIG'13) (IEEE), 1–8. doi:10.1109/cig.2013.6633618

Schuemie, M. J., Van Der Straaten, P., Krijn, M., and Van Der Mast, C. A. P. G. (2001). Research on Presence in Virtual Reality: A Survey. *CyberPsychology Behav.* 4, 183–201. doi:10.1089/109493101300117884

Scott, B. (2002). The Illusion of Intelligence," in *Ai Game Programming Wisdom*. Hingham, MA: Charles River Media, Inc., 16–20.

Shaker, N., Togelius, J., Yannakakis, G. N., Poovanna, L., Ethiraj, V. S., Johansson, S. J., et al. (2013). "The Turing Test Track of the 2012 Mario AI Championship: Entries and Evaluation," in IEEE Conference on Computational Intelligence in Games (CIG'13) (IEEE), 1–8. doi:10.1109/cig.2013.6633634

Soni, B., and Hingston, P. (2008). "Bots Trained to Play like a Human Are More Fun," in IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 363–369. doi:10.1109/ijcnn.2008.4633818

Tencé, F., Buche, C., De Loor, P., and Marc, O. (2010). "The challenge of Believability in Video Games: Definitions, Agents Models and Imitation Learning," in 2nd Asian Conference on Simulation and AI in Computer Games (GAMEON-ASIA'10). Editors W. Mao and L. Vermeersch (Belgium: Eurosis), 38–45.

Tencé, F., Gaubert, L., Soler, J., De Loor, P., and Buche, C. (2013). CHAMELEON: Online Learning for Believable Behaviors Based on Humans Imitation in Computer Games. *Comp. Animation Virtual Worlds (Cavw)* 24, 477–496.

Thawonmas, R., Murakami, S., and Sato, T. (2011). "Believable Judge Bot that Learns to Select Tactics and Judge Opponents," in IEEE Conference on Computational Intelligence and Games (CIG'11) (IEEE), 345–349. doi:10.1109/cig.2011.6032026

Thomas, F., and Johnston, O. (1981). *Disney Animation: The Illusion of Life, Vol. 6*. New York: Abbeville Press.

Togelius, J. (2016). How to Run a Successful Game-Based AI Competition. *IEEE Trans. Comput. Intell. AI Games* 8, 95–100. doi:10.1109/tciaig.2014.2365470

Togelius, J., Yannakakis, G. N., Karakovskiy, S., and Shaker, N. (2012). "Assessing Believability," in *Believable Bots: Can Computers Play like People?* Editor P. Hingston (Springer Berlin Heidelberg), 215–230. doi:10.1007/978-3-642-32323-2_9

Turing, A. M. (1950). I.-Computing Machinery and Intelligence. *Mind* LIX, 433–460. doi:10.1093/mind/lix.236.433

Van Hoorn, N., Togelius, J., Wierstra, D., and Schmidhuber, J. (2009). "Robust Player Imitation Using Multiobjective Evolution," in 2009 IEEE Congress on Evolutionary Computation (IEEE), 652–659. doi:10.1109/cec.2009.4983007

Verhagen, H., Eladhari, M. P., Johansson, M., and McCoy, J. (2013). "Social Believability in Games," in Advances in Computer Entertainment. ACE 2013. Lecture Notes in Computer Science. Editors D. Reidsma, H. Katayose, and A. Nijholt (Cham: Springer), 8253. doi:10.1007/978-3-319-03161-3_74

Yan, J., and Randell, B. (2005). "A Systematic Classification of Cheating in Online Games," in Proceedings of 4th ACM SIGCOMM Workshop on Network and System Support for Games (New York, NY, United States: ACM), 1–9. doi:10.1145/1103599.1103606

Yannakakis, G. N., and Martínez, H. P. (2015). Ratings Are Overrated!. *Front. ICT* 2, 5. doi:10.3389/fict.2015.00013

Conflict of Interest: Author CE was employed by Virtualys.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.