Bot Believability Assessment :
a Novel Protocol & Analysis of Judge Expertise

Cindy Even
Virtualys and Lab-STICC, ENIB
Email: even@enib.fr

Anne-Gwenn Bosser and Cédric Buche
Lab-STICC, ENIB
Email: {bosser,buche}@enib.fr

***Abstract*—For video game designers, being able to provide both interesting and human-like opponents is a definite benefit to the game's entertainment value. The development of such believable virtual players also known as Non-Player Characters or bots remains a challenge which has kept the research community busy for many years. However, evaluation methods vary widely which can make systems difficult to compare. The BotPrize competition has provided some highly regarded assessment methods for comparing bots' believability in a first person shooter game. It involves humans judging virtual agents competing for the most believable bot title. In this paper, we describe a system allowing us to partly automate such a competition, a novel evaluation protocol based on an early version of the BotPrize, and an analysis of the data we collected regarding human judges during a national event. We observed that the best judges were those who play video games the most often, especially games involving combat, and are used to playing against virtual players, strangers and physically present players. This result is a starting point for the design of a new generic and rigorous protocol for the evaluation of bots' believability in first person shooter games.**

***Keywords*-Autonomous agent; virtual player; believability; multi-player video games; player expertise; evaluation.**

I. Introduction

Computer games sometimes require AI controlled virtual agents (Non-Player Characters or bots) in order to advance the storyline or substitute for human players. In this paper we are interested in the later. According to Livingstone [1] "the requirement for modern computer games is not unbeatable AI, but believable AI". Indeed, unbeatable bots are more likely to be frustrating to play against. Soni et al. [2] shows that bots trained to play like human players are more enjoyable to play against and more re-playable. A bot is considered believable when it gives the illusion of being controlled by a human player [3]. Over the years, there has been heightened interest in the development of believable bots [4, section 3.6]. However, the evaluation of such systems is complex and different approaches have been used in the past [5] giving results that can not be compared [6].

One of the possible approaches is to adapt the well-known Turing test [7] which was the inspiration behind the BotPrize competition [8]. This competition has provided highly regarded assessment methods for comparing bots believability in the First Person Shooter (FPS) game Unreal Tournament 2004 (UT2004). Its design has evolved significantly over the years. In the original version, judges were set to play blindly against a bot and a human player and had to rate each of them for humanness after the match.

According to Hingston [9], while this design had proven effective, it was logistically difficult to organise and the collection and analysis of results were laborious. A new protocol made judging part of the game by providing a modified weapon which could be used to tag an opponent as a human or bot. However, this created a new game-play and triggered the emergence of non-typical behaviours (stopping for observing opponents [10] or attempting to communicate through movements and shooting patterns [11].

For the latest version of the BotPrize, a third person assessment was added. Judges evaluate the believability of players through recorded videos of matches. The inconveniences of this method is that bots should be believable from the point of view of players (not watchers) [1] and that there is no established protocol for selecting the videos [6]. In this paper we focus only on first-person assessment protocol where judges play the game.

This paper describes a novel evaluation protocol based on the first version of the BotPrize competition and a system which allows us to partially automate the execution of the competition. Both were used during a national competition, the finals of which took place during the PFIA 2017 platform[1]. We took advantage of this event to collect and analyse data on the competition and the judges' gaming habits. It allowed us to profile participants (in this paper this term refers to individuals who participated in the jury and not the competitors) based on their expertise in video games and their performance to distinguish bots from humans. We observed that the best judges were those who play video games the most often and especially games involving combat and used to playing against virtual players, strangers and physically present players.

In section II we give an in-depth description of the novel protocol and its implementation during the competition. In section III we describe the system. In section IV, we present

[1] https://pfia2017.greyc.fr

the method of the experiment. section V presents the results obtained which are discussed in section VI. We conclude with suggestions for improvement in section VII.

II. The Competition

We describe here the competition set-up. It was run in a number of rounds.

Match format: 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. Judges play against a bot or against another judge.

Match ending condition: We defined a game ending condition in order to ensure a similar number of encounters between each judge and player. Using a maximum duration or a goal score to reach were not sufficient criteria so we count the total amount of frags that occur during the match. 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). Every time a frag occurs in the game, we increase a counter and once this counter reaches the limit we set, the game ends automatically. We also set a maximum duration as a security to make sure the game does not last too long for logistical reasons.

Assessment method: We used a binary scale coupled with a certainty scale to collect the participants' judgements (the BotPrize used a Likert scale). While previous work [12] 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. Binary scales have been proven to be equally reliable, quicker and perceived as less complex [13] than traditional rating-based questionnaires. A certainty scale was added in case the participant hesitate between two proposals : according to Krosnick[14], using a third "I do not know" option instead can result in the decision not to do the cognitive work necessary.

Number of judges: In order to be able to profile the judges according to their level of expertise, we wished to involve as many participants as possible in the jury of the competition. To cater for people who had never played UT2004 before, we included a training phase. We also added a final questionnaire to get some information about the participants' gaming habits.

Our protocol can be summarised as below. We kept the presentation similar with [8] to facilitate comparison with the original BotPrize.

A) Training phase.

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.
    4) The judges were automatically connected to the game on their assigned server.
    5) Each game was a Death Match.
    6) At the end of the round, each judge was asked to fill a form to judge their opponent.
    7) After a short break, the next round starts.

C) Final questionnaire.

III. Implementation

Evaluation is subjective in nature, it is important to collect as many judgements as possible. This is only practical if this task is partly automated, which led us to the system design we describe here, composed of three modules linked together via various communication protocols (see Figure 1).

The BotContest Application: provides a user interface allowing the investigator to select the parameters of the competition and to follow its progress. It manages all the execution of the competition (starting and stopping servers, connecting players and bots, . . .), and registers game logs in the database.

The BotContest mod for UT2004: The mod enforces anonymity for bots and players (appearance, removal of statistics and chat, . . .), implements the game ending condition and records game logs.

The BotContest Website: displays information explaining the task the participants will have to perform, and collects and records their judgements in the database.

Physical arrangement: The BotPrize required two rooms, in order to separate confederates from judges. Our proposed physical arrangement now requires only five computers (one for the investigator and four for participants) in one room. Participants are set apart and use headphones to ensure they do not guess when they play against another player.

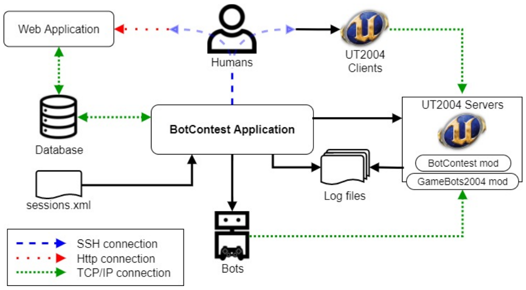


Figure 1: System architecture.

IV. Experiment Methodology

We used this competition to analyse the judges' expertise according to their video game habits. In this section we detail the characteristics of this experiment.

A. Participants

1) Competitors: The competition was open to everyone (academic, professional and independent). Six teams entered the competition out of which three qualified for the finals.

2) Judges: Everyone attending the PFIA 2017 conference was invited to take part in the jury. Over the three days, sixty members of the national artificial intelligence research community participated.

B. Procedure

Participants were welcomed in groups of four. On their arrival, a web page was already opened on each workstation with the following indications[*]:

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.

The experiment then continued with a training phase where participants were provided information about the game, its controls, weapons and power-ups. After reading this page, a 3-minute UT2004 match would automatically load. During this match, participants trained against a native bot. Once the match over, a web page would display the judging questionnaire for the opponent. This ensure their first opponent will be judged under the same conditions than their next ones.

The second phase of the experiment consisted of four rounds where the participants played a match of UT2004 with the BotContest mod and then filled the judging form. During the four rounds, the participants would face the three bots and one of the other participants.

In the final phase, participants were invited to complete a questionnaire collecting information regarding their gaming habits (see subsection IV-D for a detailed description.).

C. Independent Variables

1) The four maps: DM-1on1-Albatross, DM-1on1-Spirit, DM-1on1-Idoma and DM-Gael, were selected for their small size (more appropriated for one-on-one death-match games). The participants played on a different map for each match. This prevents them from learning the map and developing strategies that would lead them to judge opponents differently from one match to another.

2) The TimeLimit: it was set to 5 minutes making the whole experiment last approximately 30 minutes, befitting hosting conference constraints and a threshold detected during the preliminary qualification process. Indeed, it was noticed that some bots could not maintain a believable behaviour more than three minutes.

3) The FragLimit: was set to 10 after extensive testing showed it allowed to obtain a match duration closer to 5 minutes on average.

D. Measures

For each game we collect : the map used, the duration of the match, the winner of the match, the score of the two players as well as the number of times they fragged, committed suicide and killed their opponent. The judgement given by the participants after each match as well as their degree of certainty was also recorded allowing us to calculate a humanness score and a reliability score. The score increments when the player has been judged to be a human and decrements otherwise. If the given degree of certainty was 0 (i.e. "Not sure at all"), then the score remained unchanged. Only human players have a reliability score since the bots do not judge. This score is incremented when the player has correctly judged his opponent and decremented otherwise.

At the end of the experiment, participants complete a questionnaire that evaluates their expertise[*]:

1) How often do you play video games? (one answer): 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? (mutliple answers): Computer / Console / Hand-held game console / Arcade game / Other device.

3) What types of games do you play? (ranked from most to less often): A- First-Person Shooter / B- Strategy games / C- Platform games / D- Adventure, Action Games / E-RPG: Role Playing Game / F- Educational games / G-Management Games / H- Simulation games / I- Sports Games / J- Racing Games / K- MMORPG = Massively multi-player on-line role-playing game / L- Physical or sports games

4) Do you play : (one answer): Alone / With virtual players / On-line with strangers / On-line with friends or family / With physically present players.

V. Results

A. Competition Results

The humanness scores were : -0.33 for the third bot, -0.29 for the second and -0.19 for the winner. The human players obtained a score of 0.38 on average. Scores for the bots are all negative which means that none of them passed the test. A T-Test was performed with a p-value of 9.3×10^{-7} indicating a significant difference between the humanness score for humans and bots.

The bar plot in Figure 2 shows the repartition of the matches duration for each map. The duration of matches was discretised into five classes. We note that the duration of the match differs from one map to another. To validate

[*] Translated from French.

this observation a Kruskal-Wallis test was applied, with a p-value of 4.8e-15 indicating that the mean of the match duration differs significantly depending on the maps. This confirms the observations we made during the pre-tests; on some maps, the players meet their opponents much more often than on others.

The humanness score also varies according to the map but more importantly for bots than for humans (see Figure 3). The Kruskal-Wallis test gave p-values of 0.093 for bots and 0.52 for humans. Therefore, the humanness score for bots varies significantly depending on the map. However, since the duration of the match depends on the map, which means that we must consider these results with caution. The humanness score seems to vary with the duration of the matches according to the bar plot in Figure 4 : the shorter the matches, the lower the score. The Kruskal-Wallis test gave p-values of 0.39 for bots and 0.38 for humans, which does not allow us to reject the null hypothesis.

We also studied a possible link between the humanness score and (a) the fact that the player won, (b) his score and (c) the number of times he died of his own actions. The Kruskal-Wallis test gave the p-values: (a) 0.67, (b) 0.52, (c) 0.76. We can not reject the null hypotheses so there is no link between these elements and the humanness score.

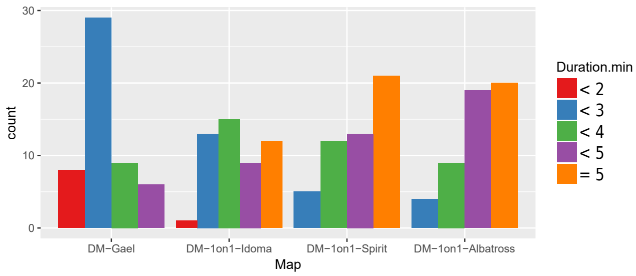


Figure 2: Bar plot of the match durations (in minutes) depending on the maps

B. Experiment Results

Using the final questionnaire we analysed four characteristics of gaming practices : 1) gaming frequency, 2) usual type of game played, 3) usual devices and 4) type of players usually faced. We profiled the participants into three expertise level groups according to their reliability scores. Many participants had an identical intermediate score so we distributed them as follows : 10 best - 40 intermediate - 10 worst. The best judges are those who have correctly identified all their opponents. The worst were wrong at least 3 times out of 4.

1) Gaming Frequency: In order to determine a possible dependency between the participants' level of expertise and their gaming frequency, we established a contingency table. The chi square of independence between the two variables is equal to 11.74 (p-value = 0.30) so we can not reject the null hypothesis. However, the result of the correspondence analysis[2] (see Figure 5a) is rather interesting : it shows that the best judges tend to play everyday, the worst tend to never play, and intermediate judges play occasionally.

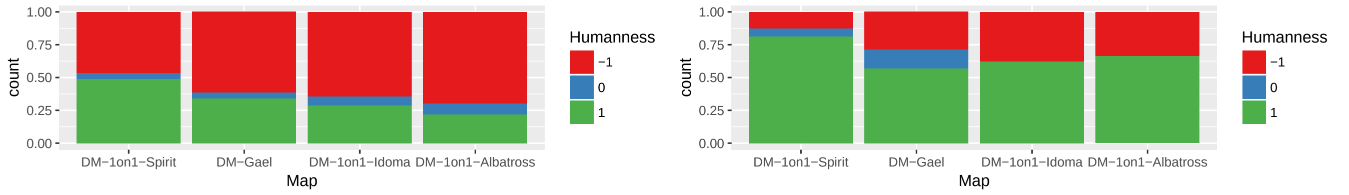
2) Usual Type of Game: Using the same method we obtained a chi square of independence between the two variables of 31.60 (p-value = 0.024). We conclude that there is a dependence between the level of expertise of the participants and the type of video game they usually play. The result of the correspondence analysis (see Figure 5b) allows us to obtain more information about this dependence. The letters in red on the figure correspond to the type of games as given in subsection IV-D. This graph allows us to see that participants with the highest level of expertise play games such as (A) first-person shooter games and (D) adventure and action games. For both these games shooting and fighting are main components. Participants with intermediate judging level play games such as (B) Strategy games, (E) Role Playing Game and (C) Platform games. In these types of game, it is quite common to encounter combat phases but they are not a main component of the game. Participants with the worst level of expertise rather play games such as (I) Sports Games, (J) Racing Games, (K) MMORPG and (G) Management Games. These types of games do not normally contain shooting phases, or at least, this is very rare.

3) Usual Devices: The distribution of the answers chosen by the participants concerning the devices used is similar for all levels of expertise. There is therefore no link between these two elements.

4) Usual player types faced: Table I shows the distribution of responses for each level of expertise. We note that the participants with the best level of expertise are those who tend to encounter all types of players unlike the other participants. To confirm these observations we performed a multiple correspondence analysis. This method locates all the categories in a Euclidean space. To examine the associations among the categories, the first two dimensions of the Euclidean space are plotted (see Figure 6). On this graph, the values 1 indicates the positive answer (i.e. the participant claimed to be used to play with this type of player), while 0 indicates the negative answer. We can see on this graph that all the positive values are on the left while the negative values are on the right. The best judges are located to the left of the graph, while the worst and intermediate ones are more to the right. This shows that the values on the right are more shared among the participants with the best level of expertise than the others and thus confirms our observations made from Table I.

[2] Greenacre, Michael (2007). Correspondence Analysis in Practice, Second Edition. London: Chapman & Hall/CRC.

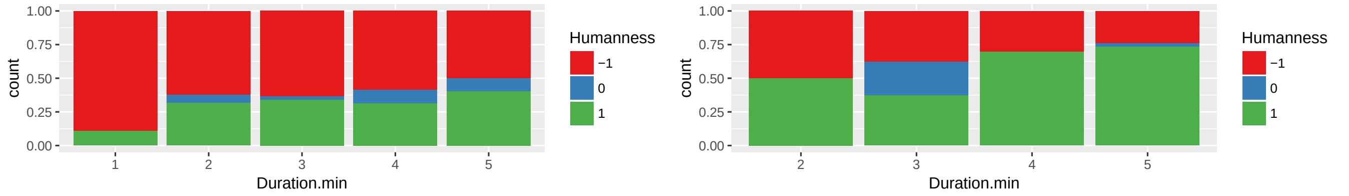
[3] Physically Present Players



(a)

(b)

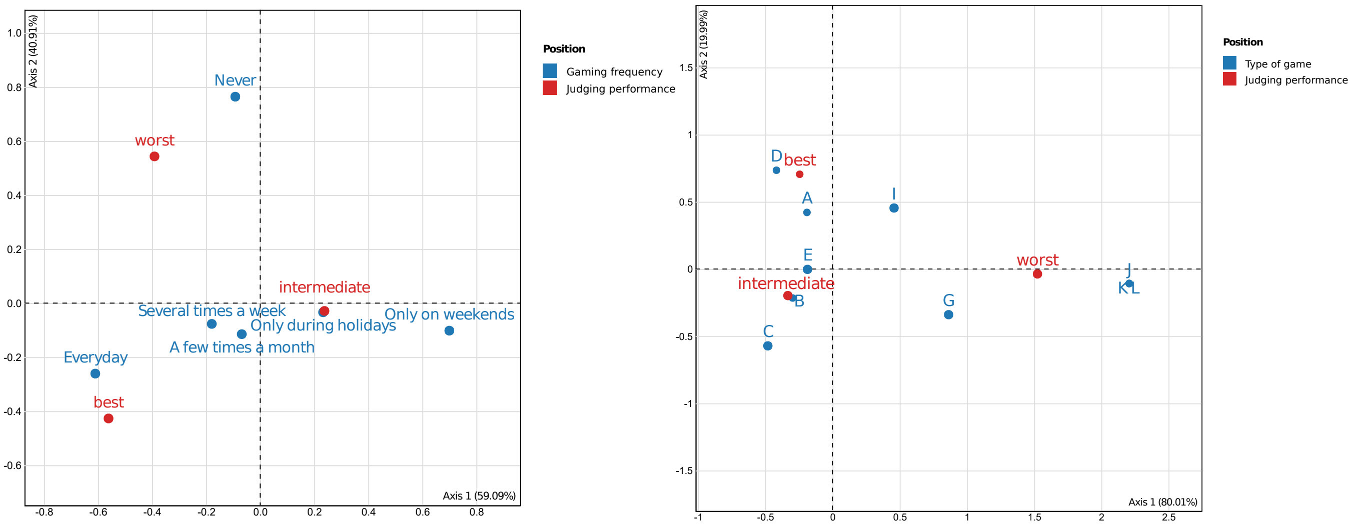
Figure 3: Bar plot of the humanness score for (a) bots and (b) humans depending on the maps.



(a)

(b)

Figure 4: Bar plot of the humanness score for (a) bots and (b) humans depending depending on the match durations (in minutes).



(a) Correspondence analysis factor map

(b) Correspondence analysis factor map

Table I: 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[3]
Best	100	60	70	70	90
Intermediate	80	33	28	48	48
Worst	70	40	40	70	40

VI. Discussion

This study allowed us to make some interesting observations both on the characteristics of the competition and on the level of expertise of the participants. Firstly, we noticed that the number of times the players meet depends on the map used for the match. Moreover, bots are perceived as being more human-like on some maps than on others : depending on the map, different behaviour may be expected. On the DM-Gael map for example the matches are fast-paced which is not surprising since it is composed of a single room where it is particularly difficult to hide. Thus, close combat is more likely to be carried out on this type of map than sniping. It therefore seems important to integrate different maps when assessing the believability of the bots, in order to observe these different strategies.

We also noticed that neither the score nor the fact that the player has won or lost has an influence on his humanness score. This is particularly interesting : player performance

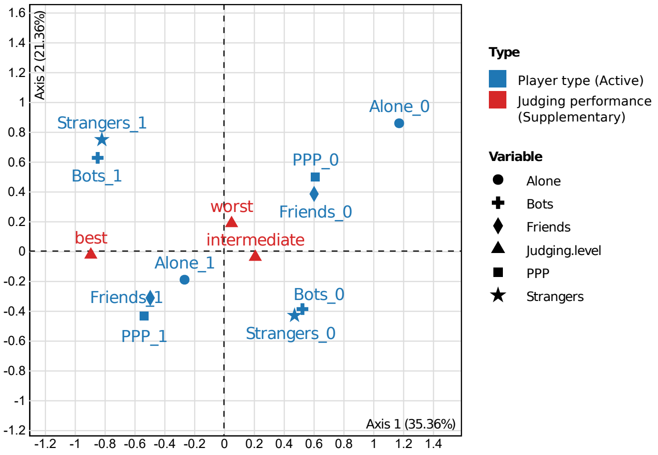


Figure 6: Multiple correspondence analysis plot for dimensions 1 and 2

and believability seem unrelated.

The results of the experiment allowed us to profile the participants with the best level of expertise for distinguishing bots from human players : 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). Participants 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 players do not sufficiently master the type of game used for the competition to have the ability to judge their opponents effectively. Even if the rules of the game are very simple (kill the opponent a maximum of times), it is nevertheless difficult to acquire the necessary skills to be able to master this type of game. Despite the addition of the training phase, we noticed that some participants, who had never played this type of game before, had difficulty even to navigate the environment. Some of these players were also surprised by certain behaviours such as opponents jumping after being seen even in the absence of obstacles. Yet this behaviour is often encountered in FPS since it is more difficult to realise a head-shot on a jumping enemy. Players will expect different behaviour depending on their expertise which illustrates very clearly the subjectivity of believability.

VII. Conclusion and Future Work

We presented a novel evaluation protocol based on an early version of the BotPrize and a system partly automating the execution of a competition. It allowed us to implement a competition during a national event and to easily involve sixty invited judges. By way of comparison, only five judges were part of the jury of the BotPrize competition.

Data gathered during the competition suggest possible improvements : the map used during matches can have an impact on the humanness score. In the current configuration, participants play on different maps for each match but they encounter a different opponent on each of them. A more rigorous protocol may present the judge with the same opponent on different maps at the cost of the evaluation duration.

Finally, despite the effort to keep the evaluation process out of the game, some judges still put in place strategies to unmask the nature of their opponents rather than play. Conducting the evaluation in the form of a competition may be the cause of these behaviours as volunteers being invited to join the jury feel unconsciously pushed to judge rather than play. In an ideal evaluation context, judges would ignore the experiment's aim.

References

[1] D. Livingstone, “Turing's test and believable AI in games,” *Computers in Entertainment*, vol. 4, no. 1, p. 6, jan 2006.

[2] B. Soni and P. Hingston, “Bots trained to play like a human are more fun,” in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, June 2008, pp. 363–369.

[3] F. Tencé, C. Buche, P. De Loor, and O. Marc, “The challenge of believability in video games: Definitions, agents models and imitation learning,” in *GAMEON-ASIA'2010*, 2010, pp. 38–45.

[4] S. M. Lucas, M. Mateas, M. Preuss, P. Spronck, and J. Togelius, “Artificial and computational intelligence in games (dagstuhl seminar 12191),” in *Dagstuhl Reports*, vol. 2, no. 5. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[5] I. Umarov and M. Mozgovoy, “Believable and effective AI agents in virtual worlds: Current state and future perspectives,” *International Journal of Gaming and Computer-Mediated Simulations*, vol. 4, no. 2, pp. 37–59, 2012.

[6] C. Even, A.-G. Bosser, and C. Buche, “Analysis of the protocols used to assess virtual players in multi-player computer games,” in *Proc. IWANN*. Cham: Springer International Publishing, 2017, pp. 657–668.

[7] A. M. Turing, “Computing machinery and intelligence,” *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[8] P. Hingston, “A turing test for computer game bots,” *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 1, no. 3, pp. 169–186, Sept 2009.

[9] ——, “A new design for a turing test for bots,” in *Proc. CIG*, Aug 2010, pp. 345–350.

[10] R. Thawonmas, S. Murakami, and T. Sato, “Believable judge bot that learns to select tactics and judge opponents,” in *IEEE Conference on Computational Intelligence and Games (CIG'11)*, 2011, pp. 345–349.

[11] M. Polceanu, “Mirrorbot: Using human-inspired mirroring behavior to pass a turing test,” in *IEEE Conference on Computational Intelligence in Games (CIG'13)*. IEEE, 2013, pp. 1–8.

[12] G. N. Yannakakis and H. P. Martínez, “Ratings are Overrated!” *Frontiers in ICT*, vol. 2, no. July, p. 5, 2015.

[13] S. Dolnicar, B. Grün, and F. Leisch, “Quick, simple and reliable: Forced binary survey questions,” *International Journal of Market Research*, vol. 53, no. 2, p. 231, 2011.

[14] J. A. Krosnick, “The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be,” *Survey nonresponse*, pp. 87–100, 2002.