# The 2014 BotPrize Humanness Assessment Method

Based on recent research we have carried out comparing first-person and third-person assessment methods in the BotPrize environment, we have concluded that obtaining high humanness ratios based on third-person judgments is usually much more complicated.

In the original BotPrize judging protocol the only source of votes (deciding whether a character is a human or a bot) came directly from human judges that were forced to be active part of the deathmatch game. This is a first-person assessment method as the judge experiences from a first-person perspective the situation and behavior he/she is actually judging. Although this is a reasonable way to test for humanness in video games, as the on-line human player is the user we want to deceive with our intelligent bots, some argue that third-person methods should be used as well. For that reason, we have decided to enhance the humanness ratio measure adding more off-line third-person judges.

In sum, we now perform the humanness ratio calculation based on two sources of votes:

* **First-Person Assessment (FPA)**: using the BotPrize in-game judging system.
* **Third-Person Assessment (TPA)**: using a crowdsourcing platform.

Therefore, the humanness of a bot is calculated as a function of both FPA and TPA measures:

***Humanness = FPA\*FPWF + TPA\*TPWF;***

FPWF 🡪First-Person Weighting Factor = 0,5.

TPWF 🡪Third-Person Weighting Factor = 0,5.

Both FPA and TPA measures are calculated taking into account the reliability of judges, in other words, votes from best judges have more impact in the final measure (see below).

Judging reliability is not calculated only by counting the number of correct judgments divided by the total number of judgments (because that could full the measure). Instead, we use the Signal Detection Theory[[1]](#footnote-1), taking into account both hits, misses and false alarm rates.

|  |  |  |
| --- | --- | --- |
| Judge SDT Matrix | **Vote “Human”** | **Vote “Bot”** |
| **Player is a Human** | *Hit* | *False Alarm* |
| **Player is a Bot** | *Miss* | *Hit* |

False alarms and misses have to be taken into account. Therefore, judge reliability (*JR*) is calculated as follows:

Where: *j* is the index of the judge, or judge identifier; *Nj* is the total number of votes issued by judge *j*.

In addition to the *JR*, which is calculated for each judge, a measure about the relative reliability (*JRR*) of each judge is calculated as follows:

Where, *AvgJR* is the arithmetic mean of *JR* measures from all judges:

Where, *J* is the total number of judges.

Once we have these measures about the reliability of judges, we calculate the humanness of characters using two different approaches (calculating *FPA* and *TPA* measures).

## Details about the First-Person Assessment method

First-Person judging will be done using the same in-game judging system that was used in 2011.

The system is based on a modification to the Link Gun. The primary mode is intended for tagging BOTs, while the secondary mode is for tagging HUMAN-controlled opponents.

During the game, when a player believes he/she has identified an opponent as a BOT (respectively a HUMAN), he/she shoots the opponent using the primary (respectively secondary) mode of the Link Gun. The shot has no effect on the opponent, but the player will see a tag "BOT" (or "HUMAN") attached to the opponent to remind them of which opponents have been judged. If the player changes his/her mind, the opponent can be shot using the other mode to reverse the judgement, as often as he/she desires.

The final tag for each opponent at the end of the round constitutes that player's judgement on that opponent. At the conclusion of all the judging rounds, each player's or bot's *humanness rating* will be the percentage of times it has been judged human over all the times it has been judged.

During the First-Person Assessment sessions, all characters will be evaluated from two perspectives:

* **As judges**, they will get a judgment relative reliability (*JRR*) ratio.
* **As players**, they will get a humanness ratio (*FPA*).

*JRR* is calculated as described above. *FPA* is calculated as follows:

Where, *FPAi* is the First-Person Assessment score for player *i*; *Ni,j* is the total number of votes issued by judge *j* to player *i*; *J* is the total number of judges; and *humannessi,j* is the humanness ratio of player *i* according to judge *j*, which is calculated as follows:

Where, *Missi,j* is the total number of times judge *j* issued a vote “Human” for bot *i*, and *Ni,j* is the total number of votes issued by judge *j* to player *i*.

Judges will be encouraged to play the game as normal, in addition to judging. As incentives for judges to do this, there will be awards for:

* Best judge (best *JRR* measure), and
* The judge with the highest FPA rating (remember all players get a FPA score, including judges).

## Details about the Third-Person Assessment method

During the Third-Person Assessment sessions, all characters will be evaluated only from one perspective:

* **As players**, they will get a humanness ratio, and therefore a TPA measure along the same lines as the FPA is calculated.

The Third-Person Assessment method is based on a crowdsourcing scheme and we will have a great number of human judges. Each of these anonymous human judges will assess the believability of several characters by watching (from a third-person perspective) selected video clips recorded during the First-Person Assessment sessions.

The reliability of anonymous human judges (*JR*) will be also calculated by our system and they will get paid according with their performance in the task of telling apart humans and bots.

The performance of third-person anonymous human judges is also calculated using the Signal Detection Theory.

### Third-Person Assessment Questionnaire

Using the crowdsroucing platform a number of anonymous judges will be presented with video clips from the first-person assessment sessions. A set of these clips is obtained for every player (selecting its specific avatar in the spectator mode) by taking the central 5 minutes from each 15 minutes deathmatch. In other words, if *P* is the total number of bots and *D* the total number of deathmatch sessions, there will be *P* x *D* video clips available for third-person assessment.

In this 2014 edition of the BotPrize we have 5 days of third-person session of 1 hour duration (each day 4 maps of 15 minutes are used). Therefore, we have 20 different deathmatches. If we have 5 competitor and 5 judges as players in these deathmatches (a total of 10 players), 200 clips will be generated (10 players x 20 deathmatches: a total of 200 x 5 minutes = 1000 minutes/16+ hours of video).

It is important to remark that in the sample of 200 clips obtained from the first-person deathmatches, 50% of players are human judges and the other 50% are AI bots.

When anonymous judges are asked to participate in the third-person assessment they are asked the following questions:

* **Before the user watches the 3 minutes video clip**
  + Age (16 – 125 years):
  + Gender: (Male | Female | T | Other).
  + Firt-Person Shooter Gaming Experience: (Rookie | Gamer | Fequent Gamer | Hardcore Gamer ).
  + How long have you been playing video games: (0 – 100 years).
  + Education: (primary school | secondary school | high school | graduate | postgraduate ).
  + Job Title: (free text).
  + **Instructions before watching the video**: “You are about to watch a 3 minutes video clip in which you will see a virtual character playing a first-person shooter video game. Your assigned task is to carefully observe the behaviour of the character, then you will be asked if you think the character is controlled by a human player or controlled by an Artificial Intelligence program. In other words, your mission is to decide whether or not the behaviour you are watching is human-like.”
* **After watching the 3 minutes video clip**
  + **Instructions after watching the video**: “Now that you have seen this character in action, what do you think? Is it a human or a machine?”
    - Is it a human or a machine controlling the virtual character? (Human/Machine).
    - To what extent you agree with these statements:
      * + The virtual character shown in the video clip is fully controlled by a human player: 0%- 100%
        + The virtual character shown in the video clip is fully controlled by an Artificial Intelligence (AI) computer program: 0% - 100%.
    - What was the primary aspect you perceived in relation to the behavior of the character that helped you most to make your decision about it being a human or a machine? [Free text]
* **Information to be recorded**
  + All data corresponding to the answers to the former questions.
  + Details of the video presented to the subject (name of the bot or player, label of the video clip).

1. <http://en.wikipedia.org/wiki/Detection_theory> [↑](#footnote-ref-1)