UNIVERSITE BRETAGNE LOIRE / MATHSTIC

ÉCOLE NATIONALE D'INGÉNIEURS DE BREST

THESE DE DOCTORAT DE

L'ECOLE NATIONALE D'INGENIEURS DE BREST
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Cindy EVEN

Proposal of a Protocol and a Computer Tool for Assessing the Believability of Virtual Players in Multiplayer Video Games

Thèse présentée et soutenue à Plouzané, le 28 Janvier 2019
Unité de recherche : Lab-STICC, CNRS UMR 6285

Rapporteurs avant soutenance :

Odile LIMPACH	Professor, Cologne Game Lab
Daniel MESTRE	Professeur des universités, Université Aix-Marseille

Composition du Jury :

Examinateurs :	Odile LIMPACH	Professor, Cologne Game Lab
	Daniel MESTRE	Professeur des universités, Université Aix-Marseille
	Ronan CHAMPAGNAT	Maître de conférences HDR, Université de La Rochelle
	Fred CHARLES	Principal lecturer, Bournemouth University
Dir. de thèse :	Cédric BUCHE	Professeur des universités, École nationale d'ingénieurs de Brest
Co-encadrante :	Anne-Gwenn BOSSER	Maître de conférences, École nationale d'ingénieurs de Brest

Invité(s)

Julien SOLER	Ingénieur de recherche, Virtualys

PREAMBLE

This thesis was carried out at the European Center for Virtual Reality (CERV) as part of a CIFRE agreement. The CIFRE (Conventions Industrielles de Formation par la Recherche) system was set up to contribute to the innovation process of French companies and to their competitiveness. It also encourages exchanges between public research laboratories and socio-economic circles.

The CIFRE agreement involves four actors:

- The PhD student who only intervenes on the research work *(here: myself, student of the Ecole National d'Ingérieurs de Brest (ENIB) and the Doctoral School Ed-Mathstic)*,
- The company that recruits the PhD student in order to entrust him with a strategic research mission *(Virtualys)*,
- The academic research laboratory which supervises the work carried out by the PhD student *(Lab-STICC, CNRS - UMR 6285)*,
- And the National Agency for Research and Technology (ANRT), which arranges the contract with the company.

Virtualys is a software engineering and R&D company specialising in virtual reality, interactive 3D and new web technologies. Founded in 1997 by the association of four newly graduated engineers and 13 associates (mainly teacher-researchers), whose will was to enhance the research work of the CERV (European Center for Virtual Reality) which is the computer science laboratory of the National School of Engineers of Brest (ENIB).

CONTENTS

List of Figures

List of Tables

ACRONYMS

AFK Away From the Keybord 72

AI Artificial Intelligence 2, 9, 13, 15, 28, 32

ECA Embodied Conversational Agent 14

FPS First Person Shooter 3, 5, 16, 19, 23, 25, 30, 33, 69, 74, 83, *Glossary:* FPS

LAN Local Area Network 5, 45

MITM Man-in-the-middle attacks 3

MMO Massively Multiplayer Online game 4, *Glossary:* MMO

MOBA Multiplayer Online Battle Arena 4, *Glossary:* MOBA

NPC Non-Player Character 4–6, 8, *Glossary:* NPC

UT2004 Unreal Tournament 2004 15–17, 23, 34, 38, 39, 41, 46, 49, 53, 58, 72, 76

GLOSSARY

chatbot or chatterbot, is a computer program designed to simulate conversation with human users, especially over the Internet. 15

FPS A First Person Shooter is a video game genre where "first person" refers to the point of view : the player assumes the field of vision of the protagonist, so that the game camera includes the character's weapon, but the rest of the character is not seen. "Shooter" refers to the action : the game interaction largely involves moving, aiming, and shooting a gun. First-person shooter games first became popular with the release of Doom in 1993. 3,

griefer (or grief player) A player who derives his/her enjoyment not from playing the game, but from performing actions that cause grief to the opponents and disrupt their enjoyment of the game. (Mulligan and Patrovsky, 2003; Rubin and Camm, 2013) 3, 7

hitbox An invisible shape commonly used in video games for real-time collision detection. 4

MMO An online game with large numbers of players, typically from hundreds to thousands, on the same server. 4,

MOBA Also known as action real-time strategy (ARTS), is a game genre where the objective for the player teams is to defend their home-base from being destroyed by the opposing team. The first team to destroy the opponent's base wins (usually a single structure in the center of the base, often referred to as an Ancient, serves as the game ender). 4,

NPC A video game character that is controlled by the game's artificial intelligence (AI) rather than by a player. It can serve a number of purposes in video games as shown in 1.1.2. 4,

overfitting in machine learning, overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. 20

CHAPTER 1

1. INTRODUCTION

Who has never played video games? This hobby is one of the favourite of French people. In 2017, the percentage of regular players in the French general population was estimated at around 53% [1]. All generations take part in this activity and the average age of French players is around 34 years old. Over the last twenty years, the industry as well as its consumer practices have evolved significantly. Video games gradually gained popularity and can now be be found in nearly every French home. One of the markers of this success is the historical turnover of the video game market which reached 4.3 billion euros in 2017 with a record growth of 18% [2].

The history of video games goes as far back as the late 1940s when the first ever electronic game was patented: the "Cathode ray tube Amusement Device" by Goldsmith Jr. and Mann. The device simulates the trajectory of a missile being fired at targets on a cathode ray tube screen. Its trajectory can be controlled by the player by adjusting buttons to reach the still targets on the screen. After that, academic computer scientists began designing simple games and simulations as part of their research. But it is not until 1971 that the first coin-operated video game was commercially sold : the *Computer Space*. A year later, the *Magnavox Odyssey* was the first home console to be commercialised. Video games were finally making their way into the homes of the general public. The golden age of arcade video games ranged from 1978 to 1982. During this time, they were very popular and could be found in many

1. SELL/GfK study "Les Français et le jeu vidéo" on a base of 1023 people aged 10 to 65, October 2017.
2. SELL data, from the GSD/GameTrack/App Annie Intelligence panels at the end of 2017.

shopping centres. Affordable home consoles were also enabling people to play games on their home TVs. Since then, the video game industry has constantly evolved to become what we know now. Video games can now be played on many platforms such as PC, home or handled consoles, phones and with virtual reality headsets.

Game playing was an area of research in Artificial Intelligence (AI) (see the list of acronyms on page ix) from its inception. An early example is Ferranti's NIMROD (Ferranti, 1951). Designed for the 1951 Festival of Britain exhibition by John Bennett and built by engineer Raymond Stuart-Williams to play the *Nim* game, a mathematical game of strategy. This is the first digital computer specifically designed to play a game. It is the first system that allows a human player to play against a digital opponent. The player would make moves by pressing buttons on a panel, with each button corresponding to a light on the machine. The computer would then run through calculations to make its move based on the player's actions. However, at that time, games were mostly implemented on discrete logic and strictly based on the competition of two players, without AI. Games that featured a single player mode with enemies started appearing in the 1970s. But it was only after the success of *Space Invaders* (1978) that the idea of digital opponents was largely popularised. A famous example is *Pac-Man* (1980) which introduced AI patterns to maze games, with the added quirk of different personalities for each enemy. Quickly after, AI was used to control other types of characters in games. *Dragon Quest IV* (1990) for instance, introduced a "Tactics" system, where the user could adjust the AI routines during battles. These characters controlled by a computer program are what we call nowadays: a *bot*, short for robot. In the next section we will describe the different types of bots for readers who are not familiar with them. This will avoid any confusion in the rest of the manuscript.

1.1 Different Types of Bots

1.1.1 Cheat Bots

Nowadays, bots tend to have bad publicity. Indeed, this term is increasingly associated with cheating in online multiplayer games. Even if it implies to break the terms of the End User License Agreement (EULA) and to risk to see their accounts suspended or banned, some players might cheat in multiplayer games for a variety of reasons (Consalvo, 2009).

Some might find a game too difficult or time-consuming and would rather use cheats to fast-forward through tedious content, areas, or gameplay. Others may also wish to acquire status or prestige, and use specific techniques or programs to speed up their progression in the game. The griefers (see glossary on page xi) on the other hand, cheat only for the fun of causing distress and anger in other players and may not necessarily be tied to actual self-advancement in the game. However, cheating can have disastrous consequences on a video game. Unlike single player cheating where the player is only affecting his/her own gaming experience, multiplayer cheating affects everyone playing in the server. It can break the game's fundamental rules, and thereby ruin the enjoyment of earnestly players, or utterly destroy the game's challenge. To tackle this issue, video games companies attempt to provide some correctives by developing anti-cheat systems, but sometimes, it is not enough and game worlds can simply be abandoned due to the rampant cheating.

Cheating techniques can be divided into two categories of underlying vulnerability: inadequacies in the system design and human vulnerability (Yan and Randell, 2009). For example, cheater can exploit bugs or loopholes in the game design and implementation. They can also tamper the code of the game client or its data and configuration files. They can also modify the client infrastructure such as the graphics driver for instance, which can be modified to display the game differently. Man-in-the-middle attacks (MITM) attacks can also be introduced, which allows to intercept and/or manipulate data in real-time while in transit from the client to the server or vice versa. Cheaters can also use this type of attack to either delay responses from their opponents or delay their own answers for their advantage. There are other cheating techniques that rely on human vulnerability and are therefore not considered to be cheat bots. Some examples include social engineering (which consists in tricking players into entering their ID and password in malicious systems), collusion of players, or escaping from the game when the player is on the point of losing.

While some types of cheats can be generic and found in any games, others depend on the game genre (Yeung and Lui, 2008). Cheats in First Person Shooters (FPS) games include *aimbots*, which allow to automatically aim at opposing players with unnatural speed and accuracy, and *wall hacks*, which allow players to see, and sometime even to click, through the scenery and therefore find opponents who are hiding nearby

(Tian et al., 2012). These cheats can be used along with a *triggerbot* that automatically shoots when an opponent appears within the aiming reticule of the player. Some cheats might also increase the size of the enemies' hitbox allowing the player to shoot next to the enemy and be detected by the game as a hit instead of a miss. Another famous usage of bots are in Multiplayer Online Battle Arenas (MOBAs), where they can instantly avoid spells from other players. In financial and monetary games, like most Massively Multiplayer Online games (MMOs), players commonly called *gold farmers*, use cheats to automatically mine hight level items in the game and resell them for real cash. They can also repeatedly defeat monsters in a specific area of the game and then resell the rewards. Gold farming can affect a game's economy by causing inflation (Jin, 2006; Heeks, 2010). They may also degrade the game experience for other players since they tend to occupy the most efficient areas of the game to gain wealth and items, forcing players to compete even more to obtain these items (Heeks, 2010).

Enforcing fairness in online gaming is very important to avoid honest players to feel deprived, lose interest, and eventually leave the game. The gaming industry therefore needs to put in place effective strategies to detect cheating. One of the techniques that is almost a standard is the use of CAPTCHA (Complete Automated Public Turing Test to Tell Computers and Humans Apart)(Yampolskiy and Govindaraju, 2008). This test allows to challenge the user with a task that most humans can pass but that current computer programs cannot. However, such a system is not sufficient when the user is is a human assisted by a program or vice versa. It is therefore necessary to have other systems such as those that analyse network traffic information (Chen, Jiang, et al., 2009; Hilaire et al., 2010). Another solution is to analyse the behaviour of the players in the game (Chen and Hong, 2007; Tian et al., 2012; Alayed et al., 2013).

1.1.2 Non-Player Characters

While cheat bots are illegitimately introduced into the game by some unscrupulous players, other bots form an integral part of the game. These bots, commonly called Non-Player Characters (NPCs), are characters controlled by the game for players to interact with, as opposed to *Player Characters* which are controlled by the human players playing the game. They may serve a number of purposes in the game (Warpefelt and Verhagen, 2017). They can be friendly towards players, like *Companions*

or *Pets* for instance, who follow and assist the player character throughout the game. Other NPCs simply have the role of fulfilling a function such as *Quest Givers* who can help advancing the story line or provide side-quests. *Vendors* are also often present in video games. They can play the role of traders or shopkeepers with whom players can buy items such as weapon, ammunition or healing potions with virtual money. Hostile characters are referred to as *Enemies* (monsters, troopers, bosses, . . .). Finally, NPCs may also supply background information (history, lore, cultural attitudes) or simply populate the environment.

As we can see, each type of NPC is usually implemented to play a specific task in the game world. Most of the time, their behaviours are scripted and automatically triggered by certain actions or dialogue with the player characters. In early video games, NPCs only had monologues with the text being displayed in dialogue box, floating text or cut-scenes. Similar to this is non-branching dialogue, where the player character can initiate conversations and respond to NPCs. More advanced video games feature interactive dialogue (Collins et al., 2016), or branching dialogue, where when talking to an NPC, the player is presented with a list of dialogue options and may choose between them. Each choice may affect the conversation, as well as the course of the game.

This type of bot is not to be confused with the second meaning of "NPC" which refers essentially to regular characters controlled by employees of the game company. These "non-players" are often distinguished from player characters by avatar appearance or other visual designation, and often serve as in-game support for new players.

1.1.3 Virtual Players

In multi-player games, a special type of bot - that we call "virtual players" - can be used in place of human players. Their role is to play the game as a human player would. They may play the role of opponent or ally against other bots and human players, either over the Internet, on a Local Area Network (LAN) or in a local session.

In FPS games from the late 90s to early 2000s, virtual players were a feature that seemed almost ubiquitous. Pretty much any games with a significant multiplayer component would have bot support (as for examples : *Red Faction* (2001), *Star Wars Jedi Knight II: Jedi Outcast* (2002),

James Bond 007: Nightfire (2002) and even platform games such as *Conker's Bad Fur Day* (2001)). So did the first four editions of *Battlefield* (2002, 2004, 2005 and 2006). This game may have been the one where bots were the most needed since its maps are way bigger than the arena shooters mentioned previously and would require between 16 to 64 players. However, since 2006, bots are relatively dismissed. Nowadays, games with a big multiplayer component are not likely to have a bot match mode. They are by no means extinct, but they are kind of a luxury feature. Developing such bots is a very expensive task. The majority of NPCs encountered in games are heavily scripted and intended only to cope with a very narrow set of circumstances that are largely in the control of the developer. They do not need to be able to adapt to what is happening in the game, they are simply performing pre-scripted actions in response to predetermined triggers. Virtual players on the other hand, need to be able to play the game as closely as possible to how any other human player would. They need to be able to use all of the abilities that are available to real players, and they need to understand and react appropriately to any situation that might arise. That requires a lot more work which means a higher cost. Also, customers having greater access to the internet, their demand for such bots was naturally decreasing as they could play with other people more easily. As a result, programming bots is no longer financially worth the cost for video game companies and they would rather focus more on other features more valuable.

However, virtual players are still very much needed. Even though internet access today is much better than it was 15 years ago, there are still areas where it is not sufficient to fully enjoy a multiplayer game session. According to a study conducted by UFC-Que Choisir (2017), no less than 7.5 million consumers (or 11.1%) can not access an offer of Internet connection with a theoretical bitrate higher than 3 Mbit/s in France. But in games where timing is key, such as first-person shooter and real-time strategy games, any sort of latency need to be avoided, whether it is due to a bad network connection or a lack of processing power. Low latency means smoother gameplay as updates of game data are performed faster between the players' clients and game server. It has been shown that network delay differences between players lead to unfairness or imbalanced game (Zander and Armitage, 2004). Disruptions due to network problems are particularly annoying for players (Oliveira and Henderson, 2003) and it can significantly affect a player's decision to leave a game prematurely (Chen, Jiang, et al., 2009). For the unlucky players

who do not have a good enough connection, virtual players allow them to still be able to play and enjoy the game they purchased.

The new player experience can also be improved thanks to virtual players. They can help to get over the steep learning curve of some games by providing a relatively safe way to learn basic mechanics without being blamed by the other players for lack of skills. It is essential not to see the majority of new players quickly lose interest in the game due to a too hostile environment (Mulligan and Patrovsky, 2003, p. 198). This is also the case for players who are not very good at the game and who do not want to spend hours training. These players can also loose interest in the game since they do not enjoy being there as simple cannon-fodder, getting dismantled by expert players. Some players also prefer to avoid servers where their gaming experience can get spoiled by griefers and their toxic behaviours. Bots can therefore extend the public of a video game.

One of the most important benefits that virtual players can bring to a game is to increase its longevity. Since the late 90s, scientists warn that we may be creating a "digital dark age" (Kuny, 1998; Brand, 1999). It refers to a lack of historical information in the digital age as a direct result of outdated file formats, software, or hardware that becomes corrupt, scarce, or inaccessible as technologies evolve and data decays. The video game industry is not immune to this phenomenon. The *Preserving Virtual Worlds* project (McDonough et al., 2010) carried out between 2007 and 2010 has allowed to develop basic standards for metadata and content representation for long-term archival storage after investigating the issues surrounding the preservation of video games and interactive fiction. While these initial efforts provided hopeful signs, work is still needed to find a permanent solution. In the meantime, some companies, such as the digital distribution platform GOG.com (formerly Good Old Games), are looking for solutions to put old games back on sale. However, even after getting an old game running again on a modern computer, it is kind of a wasted effort if it was based on online multiplayer only. At least for video games providing virtual players, the servers can be easily populated, allowing the players to play their favourite old game as it was in its heyday. This is necessary to preserve video games as it provides a snapshot of the past by demoing how the game used to be played. Nowadays, AAA game companies regularly release new titles, encouraging the fan-base to jump to the new ones and to leave

the old ones as ghost-towns where only the hardcore and nostalgic players remain. For example, despite the fact that Battlefield 4 is relatively new (released in 2013), it is already difficult to find servers to play the subsidiary game modes. Virtual players can play a major role in populating these servers. Indie games who can not rely on a massive fan-base can also make use if bots. They can make the waiting much nicer when players join empty servers hopping that others will slowly trickled until the server is full. In summary, virtual players can be seen as nurturers for when a game is taking off, stand-ins when it is going well, carers during its waning years, and ghostly companions for when nostalgic players want to play the game years later.

1.2 Believability of virtual players

Powerful anti-cheat systems, as well as entertaining and engaging NPCs and virtual players, are some of the features that make the success of a video game. However, since the gaming industry is very competitive, companies are subject to severe time constraints and are expected to deliver high quality results in a very short time. They can not necessarily take the risk of investing time and effort in developing innovative and sustainable solutions for all these components of a game. For this reason, the scientific community plays an important role in bringing new, advanced and original solutions to create value both for video game companies and their consumers.

The expectations of today's gamers have evolved with improvements in game design. They now expect truly believable and realistic gaming environments with complex stories, characters and actions. Our work focuses on the believability of virtual players (or bots) in multiplayer video games. Unlike the realism of a character where the visual aspect is extremely important, its believability on the other hand, depends on its actions and strategies. Loyall (1997) clearly illustrates the difference between these two aspects with the example of the character of the Flying Carpet in the Disney animated film Aladdin: "*It has no way of being realistic, it is a totally fantastic creature. In addition, it does not have many of the normal avenues of expression: it has no eyes, limbs nor even a head. It is only a carpet that can move. And yet, it has a definite personality with its own goals, motivations and motions.*" When it comes to virtual players in video games, they are considered believable when the players have the impression that it is controlled by another human player (Tencé, Buche, et al.,

2010). Many researchers have made the point that players enjoy a game more if they believe that their opponent is another human represented in the game by an avatar, rather than a computer-controlled player. For example, in (Weibel et al., 2008), players who were convinced that they were playing against human opponents at the video game Neverwinter Nights (an online role-playing game), reported a greater sense of immersion, engagement and flow, as well as and greater enjoyment. In (Lim and Reeves, 2010), the researchers reported that the players exhibited greater physiological arousal when the opponent was introduced as a human rather than a bot. Also, in (Soni and Hingston, 2008), bots trained using examples of human play traces were found to be more challenging and enjoyable opponents than the standard scripted bots.

Over the years, different approaches have been used for the implementation of such bots. However most of the time, these bots were either not assessed, or they were evaluated using different protocols. Yet, in order to make improvements in the development of believable bots, a generic and rigorous evaluation needs to be set up, that would allow the comparison between new systems and existing ones. According to Clark and Etzioni (2016), "*standardised tests are an effective and practical assessment of many aspects of machine intelligence, and should be part of any comprehensive measure of AI progress*". Although the evaluation of bots' performance can be performed through objective measures (comparing score or time spent to complete a level), the evaluation of bots' believability is complex due to its subjective aspect.

1.3 Contribution

The objective of this thesis is to provide a solution for assessing the believability of virtual players in multiplayer video games. The literature review allowed us to analyse the existing protocols and thus to identify seven characteristics that vary significantly from one assessment to another. Our analysis also highlighted that evaluation methods frequently modify the gameplay, introducing a significant risk of bias. This is a serious shortcoming as virtual players are thus assessed in a specific context and not in the context of the game the way it should be played: this could skew the results of the assessment. We consequently embarked on a system of trial and error, each new protocol drawing on the successes of its predecessor whilst eliminating the failures. To facilitate the implementation of these trials, we have developed a computer system that

partially automates the execution of the evaluation process. This system is flexible and has shown its genericity by being used in several configurations. Finally, we arrived at a novel proposal which allows gamers to indirectly asses the believability of virtual players by using the reporting systems traditionally used to report cheating, abuse and harassment in online video games. The goal of our proposal is to add options in reporting forms that would report the presence of bots. We hypothesised that the more often a bot is reported, the less believable it is. In order to validate our approach, we conducted an experiment which gave very promising results.

1.4 Manuscript Organisation

This manuscript is structured as follows:

Chapter 2 provides a literature review of the protocols previously used to assess the believability of virtual players. After analysing them in detail, 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.

In chapter 3 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. To facilitate the execution of the evaluation, we developed a system that partially automates the evaluation process. It's structure and implementation is also presented in detail in this chapter as well as the results of the experiment we carried out to validate our approach.

Chapter 4 presents the evaluation that we had the chance to organise during a competition that took place at the national conference: PFIA17. 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.

From the observations that we could make during our previous experiments, we came up with a completely new design, detailed in chapter 5. 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 chapter 6 we conclude with a summary of the work we have done and we provide some prospects for improvements to finalise the solution we implemented to evaluate the believability of virtual players in multiplayer video games.

CHAPTER 2

2. RELATED WORKS

Part of this chapter was published in the 14th International Work-Conference on Artificial Neural Networks proceedings (Even et al., 2017)

IWANN 2017

The popularity and sales of a video game as well as its replayability can be greatly influenced by the implementation quality of its virtual players (Scott, 2002). For example, an unbeatable bot would be frustrating to play against while a predictable one would be boring. Indeed, according to Daniel Livingstone (2006), modern video games do not require unbeatable AI but believable AI. Also, recent experimental results (Soni and Hingston, 2008) show that believable bots increase user's enjoyment. Different approaches have been adopted for the development of believable bots, such as systems based on connectionist models (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, Gaubert, et al., 2013) - to mention just a few. Generally, the proposed systems are not assessed, and when they are, the results obtained can not be compared as different protocols have been used. However, in order to make advances in this field, 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 AI in games research has been identified as one of the main challenges in game AI research (Lucas et al., 2012). In this chapter, we review evaluation techniques for assessing the believability of virtual

players and we provide a comprehensive analysis of the evaluation features. We conclude by suggesting prospects for improvement.

2.1 Defining Believability

The first notions of believability came from the character arts (Verhagen et al., 2013). According to Bates (1994), the notion of believable characters does not refer to an honest or reliable character, but one that provides the illusion of life and thus permits the audience's suspension of disbelief [1]. To create this illusion, the authors Thomas and Johnston (1981) presented the 12 basic principles of animation in their reference book on Disney animation, dealing for instance with the basic laws of physics, emotional timing and character appeal.

While no interactions are possible in animation, video games and virtual environements allow users to interact with their characters and inhabitants. The impact of these virtual agents on users has been studied for many years. The notions of presence (Schuemie et al., 2001) (the psychological sense of "being there" in the environment) and co-presence (Goffman, 1963) or social presence (Heeter, 1992) (the perception and feeling of "being with" others) are often evaluated and can be measured using self-report or behavioural measures (Bailenson et al., 2004). When the agents have the ability to generate gesture, facial expression and speech to enable face-to-face communication with users, they are called Embodied Conversational Agent (ECA). One aspect of these agents that is frequently adressed is their believabilty (Magnenat-Thalmann et al., 2005; Bosse and Zwanenburg, 2009; Bevacqua et al., 2014) which is determined by many 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 (Julian Togelius et al., 2012): *character believability* and *player believability*. Character believability refers to the belief that a character is real. Therefore, most aspects of animation and graphics rendering are very important. 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

1. term coined by Coleridge (1817), who suggested that even though a reader or a spectator knows that the story and the characters are not real, he/she may set aside his scepticism and have feelings and reactions as if it was real.

controlled by a human player (Tencé, Buche, et al., 2010) and that 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.

As we can see, the notion of believability is largely domain dependant. Our work will chiefly apply to player believability rather than character believability. For more information on character believability; refer to (Loyall, 1997; Bogdanovych et al., 2016) and the Social Believability in Games Workshop (Verhagen et al., 2013). The next section offers a review of existing methods used to assess believability of virtual players.

2.2 Assessing Believability

The Turing test is widely considered as being a pioneering landmark for believability assessment (Marcus et al., 2016). Developed by Turing in 1950, it tests the ability of a chatbot to exhibit intelligent behaviour, indistinguishable from that of a human. In its standard interpretation, a human judge converses via text-only with a human confederate and a computer program. If by using only the responses to written questions, the judge can not reliably tell the chatbot from the human, it is said to have passed the test.

A way of evaluating AI is to organise competitions. According to Togelius (2016), the advantage of competitions is that they provide fair, transparent and reusable benchmarks. Most competitions in computer games are aimed at the development of superhuman-level opponents such as the famous chess program *Deep Blue* (M. Campbell et al., 2002) or the recent *Go* program *AlphaGo* (Silver et al., 2016). In recent years we have seen the emergence of competitions oriented toward the implementation of human-like (or believable) opponents such as the 2K Botprize competition (Philip Hingston, 2009) or the Turing Test track of the Mario AI Championship (Shaker et al., 2013).

The BotPrize is particularly interesting as it has evolved significantly over the years. It was held annually between 2008 and 2014 (except in 2013) at the IEEE Conference on Computational Intelligence and Games. It is a variant of the Turing test (Turing, 1950) which uses the “Deathmatch" game-type mode of the video game *Unreal Tournament 2004 (UT2004)* developed by Epic Games, a FPS whose objective is to kill as many opponents as possible in a given time (and to be killed as few

times as possible). The different versions of the BotPrize are described below:

First version: Its first two editions were held in 2008 and 2009 (Philip Hingston, 2009) and used the same protocol (as illustrated in Figure 2.1a). They were run in 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 (from "*1: This player is a not very human-like bot*", to "*5: This player is human*"), and to record their observations. In order to pass the test, a candidate (by candidate we refer to the entity being evaluated e.g. a bot or a human player) was required to be rated 5 (this player is human) from four of the five judges.

Logistically this competition was quite difficult to implement. There were two rooms: one with a computer for each server and for each confederate, and another room with a computer for each judge. No communication between the two rooms was possible other than by the organizers, or via game play. Spectators were able to come to the judges' room to watch the games in progress.

Second version: In 2010, a new design was implemented (see Figure 2.1b) (Philip Hingston, 2010), born from the desire to make the judging process part of the game. The organisation of this new version was much more simple since there was no need for confederates or a secret room. Only one server was running continuously, where human judges and bots could connect at any time. A weapon of the game (the Link Gun) was modified for the judging process. This weapon had two firing modes (one for each button of the mouse) that could be used to tag an opponent as being human or bot (the vote was final). 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. Whereas before, players would move quickly in order to not present an easy target, in the new competition human players are easily spotted as they are tempted to stop and observe their opponents to make a judgement (Thawonmas et al., 2011). Furthermore, judges may be inclined to attempt to commu-

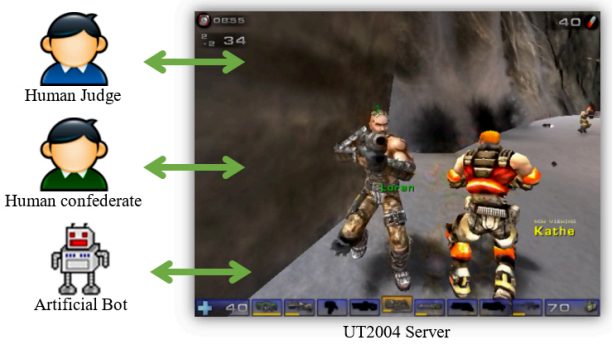
nicate through movements and shooting patterns (Polceanu, 2013). This kind of behaviour would not naturally occur in normal gameplay.

Third version: No major changes were made for the 2012 edition. The only differences were that the judges would not die if they made a wrong judgement and that 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.

Fourth version: In 2014, a consortium of Spanish researchers launched a new competition [2]. Before this version, the testing protocol of the BotPrize competition was always a first-person observation approach (i.e. the judges play the game). 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 (see Figure 2.1c). Different opinions emerge when it comes to chose whether the assessment should be first or third person oriented or a combination of both as it has been done here. More details are given in section 2.3.2.

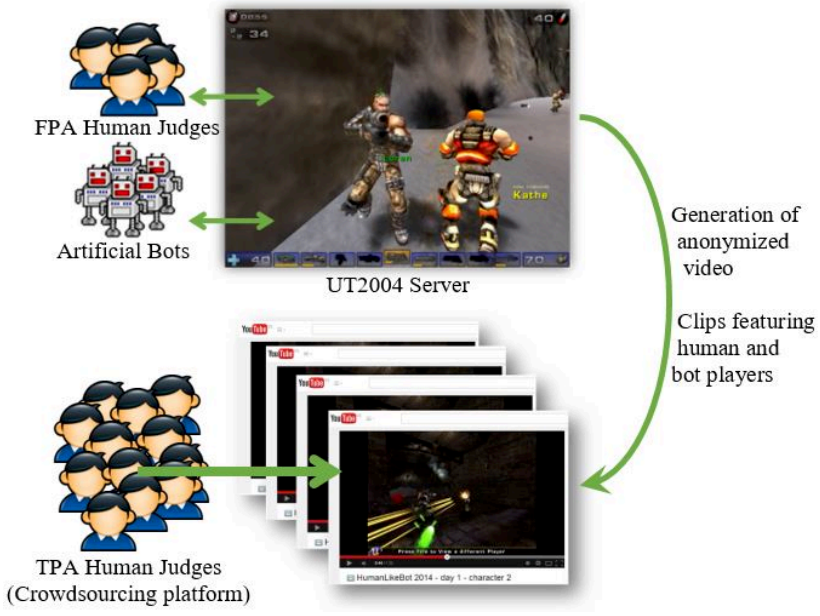
The 2K BotPrize contest was not the only protocol for assessing the believability of virtual players. Another gametype of UT2004 called “Capture the Flag" was used in (Acampora et al., 2012). The objective of the game is to capture the enemies’ flag and to return it to your own team’s flag. The authors 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

2. Human-Like Bots Competition, presented at the IEEE CIG conference by Raúl Arrabales : http://fr.slideshare.net/array2001/arrabales-bot-prize2014v2?from_action=save



(a) 1st version

(b) 2nd and 3rd versions



(c) 4th version

Figure 2.1 – Illustration of the protocol used for the different versions of the BotPrize.

and human players with different levels (novice, medium and high). After watching 4 videos, judges were asked to evaluate human-likeness on a 7-point Likert scale. An approach similar to this subjective assessment was used in (Laird and Duchi, 2001) and (Gorman et al., 2006) but with a different FPS game called Quake II. The protocol's characteristics of these player believability assessments can be found in Table 2.2 along with other relevant references of player believability (in white) and character believability (in grey) assessments.

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 (see Table 2.1): 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 fulfils 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. With this framework they evaluated the bots from three video games: *Doom* (1993), *Quake II* (1996) and *Tom GlancyâĂŹs Ghost Recon* (2001) who respectively obtained the following score: 9, 25, 42. This chronological increase in the score illustrates the effort that has been invested to improve the believability of bots over the years. A much more detailed solution was offered with ConsScale (Raul 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 FPS game bots (Raúl 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.

Tencé and Buche (2008) proposed to compute vectors (called "signatures") that characterize humans' and bots' behaviours. After defining the signatures, humans and bots are monitored in order to compute their own signatures. The distance between each bot's signature and humans' signature is calculated. The smaller the distance, the more human-like the bot. In the example provided in their paper, only two simple types of signature were used characterizing the movements in the environment.

Table 2.1 – Framework for evaluating the believability of characters from (Hinkkanen et al., 2008)

(a) Scores for movement and animation

Requirement for NPC	Points
NPC can find the most suitable path for its destination.	1
NPC's movement is not limited to a certain area, such as one room.	1
NPC's movement is not clumsy or angular.	2
NPCs are aware of each other and do not collide with each other.	1
NPC can avoid any dynamic or static obstacle in game field.	2
NPC has different animations for one action.	1
Shifting from one animation to another is fluent.	1
NPC's appearance is done carefully and no unnatural features can be found in it.	1
Total	**10**

(b) Scores for NPC's behavior

Requirement for NPC	Points
NPC makes intentional mistakes.	2
NPC has human-like reaction times.	2
NPC behaves unpredictably.	1
NPCs are aware of each other.	2
Cheating in a manner that player can not detect it.	1
Bad aim when seeing player for the first time.	1
Logical and human behavior.	1
Total	**10**

More complex signatures covering a greater range of behaviours would be necessary before considering using this solution. Also, the risk of overfitting should be taken into account when implementing such a solution since it would have a serious impact on the validity of the evaluation.

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. But unlike the solutions based on the Turing test, they do not allow to step back and evaluate the bot's behaviour as a whole. For this reason we focused on solutions based on the Turing test for the rest of the study.

As we can see from the descriptions below, the protocols used in the past for the assessment of virtual 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; Daniel Livingstone, 2006) as it depends on the per-

ceptions 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 the next section we propose to analyse characteristics of the protocols collected in Table 2.2.

2.3 Assessment's Characteristics Analysis

2.3.1 Application

The application used for the evaluation process can be pre-existing or developed specially for the test. The implementation of a sample game can be necessary when no open-source games are available (Bernacchia and Hoshino, 2014) but it needs to be well-thought-out in order to not introduce bias unintentionally. A good example from the domain of character believability is Mac Namee's simulation of a bar (2004). Two virtual bars populated by autonomous agents who could buy/drink beer, talk to friends, or go to the toilet were used. In the first simulation, the agents had long-term goals, whilst in the second they selected a new goal randomly every time they completed an action. Mac Namee noticed a difference in the results probably due to cultural effects : for the Italian subject, the random selection seemed more believable as for him as it was unrealistic to have agents returning to sit at the same table time after time, whereas for the other subjects (from Ireland), this behaviour seemed more believable. A bar environment was not necessarily an ideal choice for the evaluation as subjects had diverse expectations as to how a human would behave. This problem of cultural difference is well known to researchers interested in the development of virtual agents. Many approaches have been proposed to design agents that can adapt their behaviour to the cultural context to which they apply (De Rosis et al., 2004; Rehm et al., 2007; Lugrin et al., 2017). However, research on understanding the cultural nuances of game players is lacking (Chakraborty and Norcio, 2009). Lee and Wohn (2012) showed that there is a small effect of culture on behaviours in social network games. They found that cultural orientations affect people's expected outcomes (social interaction, recognition, relax, or relieve boredom), which in turn affects different

Table 2.2 – Comparison of the existing experiments

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 x 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 & 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 x 3 videos 1st person view	20 sec	20	✓	✓	✓	A			✓ 1 to 5	✓	online
Bossard et al., 2009	CoPeFoot	1st		n/a	48	✓		✓	C	✓				pen & paper
Philip Hingston, 2009 (BotPrize v1)	UT2004 Deathmatch	1st		10 min	5		✓	✓	A			✓ 1 to 5	✓	n/a
Philip 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 x 1 video 3rd person view	1 min	12		✓		n/a	✓				crowdsourcing platform
Acampora et al., 2012	UT2004 Capture The Flag	3rd	1 x 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 x 2 videos 1st person view	n/a	43	n/a			B		✓ 3 choices	✓ 1 to 5		online

Character believability assessment.

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.

usage patterns (giving and offering gifts to game friends, advancing in the game, customizing their avatar, publishing game status, ...). Further research on video games are required to examine cultures' effects on motives and behaviour in the virtual world (Jackson and Wang, 2013). Moreover, for this study we can investigate the French population only so we decided not to focus on the possible effects that cultural differences could have on the evaluation of bots' believability.

Choosing a pre-existing video game brings many benefits. According to Tencé, Buche, et al. (2010), it should be a multi-player game (indeed, the role of virtual players is to be played against) offering a lot of interaction between the players. Action, role playing, adventure and sport games meet these criteria. Adventure and sport games tend to be difficult to modify and in particular, they rarely offer the possibility to add customised bots. The main draw-back of role playing games is that they rely in large part on communication and natural language which is not what we intend to evaluate here. Similarly, in order to not impact the assessment, all "chat" options should be disabled (Philip Hingston, 2009). Action games, especially FPSs, are often a good choice. For the BotPrize contest, Philip Hingston (2009) chose UT2004 because it is affordable, readily available, customizable, bots and humans can play together and do not need to be collocated, and it is easy to interface a bot to the game. Julian Togelius et al. (2012) argued that FPS are not suitable for believability assessments as players encounter their opponents for only a few seconds and in the middle of a chaotic situation. For this reason they preferred to use the single player game *Infinite Mario Bros* which does not meet the criteria of being a multi-player game.

Even when using a pre-existing game, the choice of the map is very important. In the 2014 BotPrize edition, one of the maps had low gravity which affected the behaviour of the participants and where even the human players tended to exhibit bot-like behaviours (Polceanu et al., 2016).

2.3.2 1st or 3rd person assessment

Believability assessment may consider both first person and third person reports. In first person assessment, the judge has two simultaneous roles: to play the game, and to judge opponents. On the other hand, in third person assessment, the judge is only a spectator observing the game being played.

With a single player game such as *Infinite Mario Bros*, it is not possible to play the game with the bot, only a third person assessment is possible which shows that the application used for the experiment may restrict the choice of some parameters and therefore, needs to be sensible.

In (Laird and Duchi, 2001; Shaker et al., 2013; Llargues Asensio et al., 2014) the authors argued that assessing believability from a first-person perspective might be distracting since the judge has to pay attention both to the game experience and to the behaviour of the other players for the evaluation. Daniel Livingstone answered in his paper (2006) with : "*in game development the aim is to satisfy the needs of the players of a game and not those of watchers*". However, even if computer games are primarily designed for the players, video game spectating has recently become a popular activity (Cheung and J. Huang, 2011; Kaytoue et al., 2012). In Cheung and J. Huang paper (2011), the authors report that there are some spectators that actually prefer to watch professionals playing rather than playing the game themselves. On the other hand, when playing a video game, the player participates actively, unlike the spectator who can only interact with the game through communication with the player (Sjöblom and Hamari, 2017). Therefore, the direct interactions with the virtual players can only be done by the players and not the viewers.

First person assessment is possible only with applications that can be played by at least two players simultaneously. The third person assessment however, can be used with any application. When performing a third person assessment, judges are asked to give their judgement after watching a video of the game previously recorded. To reduce the subjectivity and the guesswork, Gorman et al. (2006) suggested to show more than one video to the judges in order for them to have a basis for comparison. They also pointed out the risk of introducing a bias when selecting videos for the assessment. The person in charge of the selection might pick parts of the video that could influence the responses.

When recording the videos, different points of view can be used. In some cases the application does not offer many possibilities. The Pong game for example, can only be played with a global view, representing the tennis table and the two paddles. In other video games such as FPSs, it is possible to choose between the first and third person view. Therefore,

videos can be recorded from the confederate's or the candidate's first or third person view.

Confederate's 1st or 3rd person view: The confederate's 1st person view is most commonly used for assessing the believability of bots. This might be due to the fact that it is easily recorded during game play, particularly during a first person assessment. These points of view allow us to capture the game as if the judges were in-play. The main drawback of these points of view is that a considerable portion of recording can not be used. Indeed, all the moments when the confederate is in the environment without facing the candidate are useless and need to be cut from the video.

Candidate's 1st person view: When using the candidate's first person view, the judges have less resources to evaluate the entity: for instance, they can not see its movements.

Candidate's 3rd person view: This solution has never been used in our knowledge. Yet it could be especially interesting since it would capture both the perception and the actions of the candidates. This could allow a better understanding of the decisions made by the candidate. Moreover, it would not require cuts in the recording as even the time when the candidates are alone in the environment could be used for the judgement, which would be time saving and would reduce the risk of introducing the aforementioned bias when selecting videos for the assessment.

2.3.3 Duration

The duration of video and game play varies greatly from one experience to another, going from 20 seconds to as long as the judge desires. It might depend on the nature of the game but most of the time, the choice of the experiment's duration relies on the organisers' opinion (Julian Togelius et al., 2012; Shaker et al., 2013) and is never justified. In their experiment, Soni and Hingston (2008) tried to examine the role of predictability by using two different bots during their assessment. One was deterministic and the other one was making stochastic choices which made it unpredictable. They hoped that the results would show that the second bot would be more believable but during the experiment, the subjects did not notice any difference between the two bots. The authors hypothesised that the experiment was too short and that longer sessions

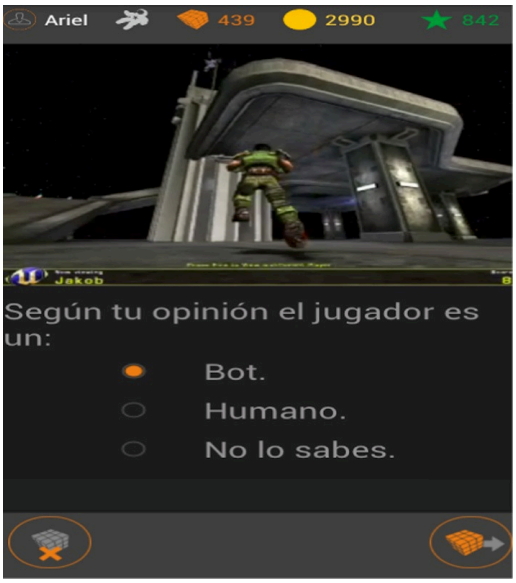


Figure 2.2 – Screenshot of the Knowxel mobile application used for the judging process in (Llargues Asensio et al., 2014) - *Translation: In your opinion, the player is a: Bot. Human. You do not know*.

could give the judges enough time to make a distinction. The observation of Paritosh and Marcus (2016) regarding the Loebner competition [3] (the first formal instantiation of a Turing Test) is similar. They argue that the test is too short (only few minutes) to allow any depth in the judgement. Even if it is important to allow enough time for the judges to make a judgement, the assessment can not be too long as it can induce inattention or mistakes due to judges' boredom or fatigue (Brace, 2008).

2.3.4 Number of judges

The assessment being of subjective nature, it seems important to collect a significantly large number of judgements in order to cancel out the biases introduced by that type of assessment (Hyman and Center, 1954). The use of online surveys eases the collection and treatment of results. For their experiment, Llargues Asensio et al. (2014) used a crowdsourcing platform for mobile devices (see Figure 2.2) that allows to conduct a video-based poll experiment where the users can vote at the end of each video clip.

3. http://www.loebner.net/Prizef/loebner-prize.html

2.3.5 Judges' and confederates' expertise

The level of the judges is sometimes taken into account for the experiment. As it has been noticed by Mac Namee (2004), the experience of players in video games can introduce a difference between the subjects. In general, for an experienced player it will be quicker and easier to recognise a bot than for a novice player. For example, in Laird and Duchi's paper (2001), only the expert player made no mistake in differentiating between bots and humans. Regarding novice players, they might not fully know the rules of the game or the set of actions available to the players which could make the whole experience too confusing and they would not be capable to sensibly evaluate the players' behaviours (Daniel Livingstone, 2006).

Another interesting element that has been taken into account in (Laird and Duchi, 2001; Acampora et al., 2012; Shaker et al., 2013) is the level of the confederates. They have a major role in the assessment as their behaviours directly influence the judges' evaluation. For example, an expert-player confederate with high performance could easily be mistaken for a bot by non-expert players who are judging (Polceanu, 2013). On the contrary, novice-players confederates who are still learning how to play the game and how to use the controls might have behaviours that could be confused with a weak bot by expert players. Confederates should be provided with sufficient time for gaining control over the game rules and commands before starting the evaluation. Philip Hingston (2009) avoided these potential problems by choosing confederates who were all reasonable level of experience, i.e. neither expert nor novice.

2.3.6 Information given to the judges

As we can see in Table 2.2, judges can be given different information before starting the experiment. Most of the time, they are informed that they will see a combination of bots and human players (A in Table 2.2). In other cases, (B) they know the nature of the entity they are evaluating. Finally, (C) judges are not informed as to the purpose of the experiment. For instance, in (Bossard et al., 2009), judges were invited to play a football video game, where all the players had a number. After a given time, the game was paused and they were given a table and the following instructions: "Cross the box corresponding to the two players controlled by humans in the simulation, if and only if you are confident in your answer.

If in doubt, write nothing". The analysis of the results revealed that judges were considerably better at distinguishing bots from human players after the first attempt.

In two other experiments, half of the participants were informed that the other character in the game would be controlled by another person, while the other half were informed that it would be controlled by a computer (AI). In fact, for all the participants, the character were controlled by a computer in (Weibel et al., 2008), and by a human in (Lim and Reeves, 2010). In the first experiment, the participants who played against the character that they believed to be human-controlled, reported stronger experiences of presence, flow, and enjoyment. And in the second experiment, the participants exhibited greater physiological arousal and reported greater presence and likeability when the character was introduced as being human controlled rather than computer controlled. These results demonstrate that the information given to the judges can significantly alter their judgement.

2.3.7 Subjective assessment types

When assessing players' believability in a game, players are asked to give their opinion (Julian Togelius et al., 2012). Their answer can have the form of a free response or of forced data retrieved through questionnaires.

Free response answers can contain much richer information but they are also much harder to analyse appropriately. Sometimes judges have the opportunity to give a free response in the form of comments (Philip Hingston, 2009). These comments can be useful for identifying areas for improvement for the bots implementation but are generally not used for evaluation.

On the other hand, by using a questionnaire, subjects are constrained to choose between some specific items, yielding data that is easier to analyse. Different types of forced questionnaires can be identified (Julian Togelius et al., 2012) :

Binary: Subjects can answer by *Yes* or *No* to a simple question (e.g. *is this player a bot?*, or, *is this bot believable?*).

Scale: Judges are asked to rate the humanness of the players' behaviour or to choose an answer within a list (e.g. *1: Human, 2: Probably Human, 3: Don't Know, 4: Probably Artificial, 5: Artificial* (Gorman et al., 2006)).

Comparison: Subjects are asked to compare two or more players (e.g. *did player A or B act more like a human player?*).

With ranking questionnaires, it is not possible to analyse the interpretation of the rating categories across subjects (Friedman and Amoo, 1999). To minimise the subjective notion of scaling and allow a fairer comparison between the subjects' answers, comparison and boolean questions can be used (Julian Togelius et al., 2012). But as mentioned by Philip Hingston (2009), a binary choice might have the effect of forcing the subjects to "toss a coin" if they are unable to choose an answer. In an effort to reduce subjectivity, in (Mac Namee, 2004; Daniel Livingstone, 2006) subjects were not asked to rate believability, instead, they were asked to compare two players and say which was more believable or acted more like a human player. The choice items may be presented in different ways, for instance, the subjects can choose between 2 solutions (*player A* or *player B*). They can also be offered more options such as *there is no difference*, or *both equally* and *none of them*, following the 4 alternative forced choice (4-AFC) protocol proposed by Yannakakis and Hallam (2009).

2.4 Discussion

When studying the protocols used in the past to assess virtual players' believability, we identified some characteristics that varied significantly from one assessment to another, giving results that can not be correlated.

2.4.1 Application

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 virtual 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.

2.4.2 1st or 3rd person assessment

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.

2.4.3 Duration

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.

2.4.4 Number of judges

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.

2.4.5 Judges' and confederates' expertise

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.

2.4.6 Information given to the judges

As we saw in 2.3.6, recent experiments have shown the influence of the information given to the judges on their judgement. This part of the assessment protocol needs to be carefully designed in order to avoid introducing a bias. When conducting a first person assessment, the gameplay might be modified if the judges know the aim of the assessment. The only way to avoid this is to keep the question secret and to ask the player only at the end of the game, whether he thought he was playing against a human player or a bot. Of course, the player could be asked only once. During a third person assessment, the best solution seems to be keeping the nature of the candidate secret and telling the judges that they would see a mix of bots and human players, so that they have no prejudices.

2.4.7 Subjective assessment types

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.

2.5 Conclusion

Virtual players play a major role in the success of video games. A new challenge is to develop believable bots that could blend in among human players. Over the years, different approaches have been used for the implementation of such bots. However, most of the time, these bots were either not evaluated, or they were evaluated using different protocols. Yet, in order to make improvements in the development of believable bots, a generic and rigorous evaluation needs to be set up, that would allow the comparison between new systems and existing ones. According to Clark and Etzioni (2016), "*standardised tests are an effective and practical assessment of many aspects of machine intelligence, and should be part of any comprehensive measure of AI progress*". Although the evaluation of bots' performance can be performed through objective measures (comparing score or time spent to complete a level), the evaluation of bots' believability is complex due to its subjective aspect.

In this chapter we analysed the protocols previously used to assess the believability of virtual 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. After an in-depth analysis of these protocols, we gave recommendations for the features that are well established. In order for the protocol to be rigorous and reusable, other features still need further study and testing to be determined.

CHAPTER 3

3. BLINDING THE JUDGES

We showed during our literature review that there are two ways to perform a Turing test for bot, using a first or third person assessment. However, direct interactions between the player and the bot are possible only with first person assessments. We have therefore favoured this type of evaluation since, in this study, we decided to focus on the experience of the players and not on that of the spectators. Our analysis has also highlighted that the main drawback with first person assessment is that it frequently adversely affects the gameplay, introducing a significant risk of bias. For instance, when players are asked to assess their opponents, they are tempted to stop and observe them to make a judgement (Thawonmas et al., 2011). But for FPS especially, a good player is always moving quickly in order to not present an easy target. Furthermore, judges may be inclined to attempt to communicate through movements and shooting patterns to unmask opponents (Polceanu, 2013). This kind of behaviour would not naturally occur during normal gameplay.

To date, there is no protocol that is rigorous and easy to implement to assess the believability of bots in video games. One of the main reason being that the gameplay of the game is affected by the evaluation process. To overcome this problem, we propose in this chapter an innovative protocol that applies a technique frequently used in scientific experiments to reduce the risk of bias wich is called "blinding". Blinding, in research, refers to a practice where study participants are prevented from knowing certain information that may somehow influence them and thereby tainting the experiment results (MacLean and Dror, 2016). Psychological research has shown how context, motivation, expectations, and experience affect people's perceptions and cognitions (Gilovich et al., 2002; Koehler and Harvey, 2004). Therefore, we decided to keep secret the real

objective of the 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.

Another problem encountered when assessing the believability of bots is the difficulty of implementing such an evaluation. Indeed, it is necessary to launch several matches of the game and to correctly connect the judges and the bots, which requires many laborious manipulations. To facilitate this process we have developed a computer system that partially automates these tasks.

Thus, in this chapter we propose a complete solution composed of an innovative protocol as well as a computer system to partially automate its execution. First, we will introduce the new model we have put in place to assess bots' believability. Then, we will detail the implementation of our system. Finally, we will describe the experiment we conducted to validate our model as well as the results we obtained.

3.1 Model

To set up our model, we were inspired by the first version of the BotPrize competition (Philip Hingston, 2009). As a reminder, this competition uses a first-person assessment method with the original version of the video game Unreal Tournament 2004 (UT2004). Indeed, it is only for the following editions that one of the weapons of the game was modified for the judgement, having as consequence to modify the main goal of the game.

We decided to run the assessment in a number of rounds, similar to the format of the first version of the BotPrize competition (Philip Hingston, 2009). However, some changes have been made and are listed below:

Information given To avoid revealing the purpose of the experiment, participants are simply informed that the they would take part in an experiment about video games. Obviously, this will not be possible in the context of a competition like the BotPrize where the objective of the competition would have been announced even before the event.

Training phase In order to familiarise the participants with the game, we added a training phase. It consists of providing information about the game, its controls, weapons and power-ups. Then, participants play a three-minutes 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.

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. Confederates are no longer necessary, instead, in each match a judge will play against a bot or against another judge.

Assessment method 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. As you can see in Figure 3.1 (see Appendix B for the original version), it 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 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 (Philip Hingston, 2009), we used a binary scale coupled with a certainty scale. While previous work (Yannakakis and Martínez, 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(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".

Final questionnaire 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.

Because of these modifications, we had to adapt the protocol of the original BotPrize. We kept the presentation similar with (Philip Hingston, 2009) to facilitate the comparison.

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 the questionnaire.

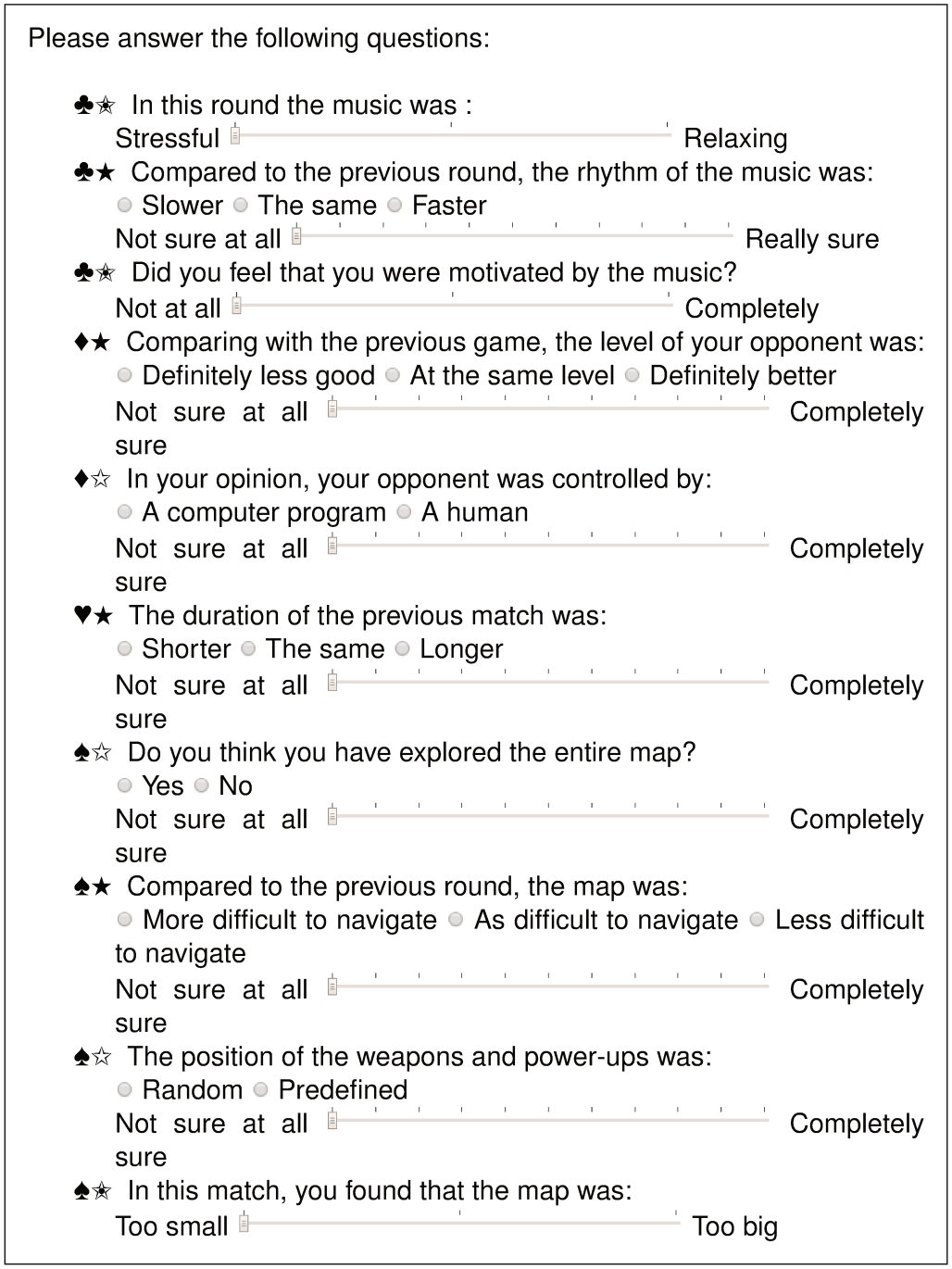
   7) After a short break, the next round starts.

C) Final questionnaire.

3.2 Implementation

According to Philip Hingston (2010), while the first version of the competition had proven effective, it was logistically difficult to organise and the collection and analysis of results was laborious. He then proposed the new design (Philip Hingston, 2010) with the modification of a weapon, making the judging process part of the game. Therefore, to facilitate the

Figure 3.1 – Questionnaire translated from French



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 3.2).

We decided to use the same video game as for the BotPrize competition: Unreal Tournament 2004 (UT2004). 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. We chose to use this game for several reasons. First, one of its big advantage is that everything (except the 3D engine) is open source and can be modified by the user which makes the game easily customisable. Then, many resources are available to ease the development of bots. GameBots (Bída et al., 2012) for instance, is a modification of the game (a mod), that provides a network text protocol for connecting to the game server and controlling in-game avatars (bots). User can control bots with text commands and receive information about the game environment. In addition to GameBots, Pogamut (Gemrot et al., 2009) provides a Java API and GUI (NetBeans plugin) that simplifies the development and debugging of the bots. Finally, we had access to resources compatible with this video game such as the executable of the winner of the last BotPrize competition: MirrorBot (Polceanu, 2013). We tried to use the latest edition of the game: Unreal Tournament 4 [1]. Unfortunately, the game is still in pre-alpha version which means it is in an early stage of development and it is not stable enough to be used in this context. Once the game is released we think it can easily be used with our model since it is very similar to UT2004.

3.2.1 UtBotEval Application

The `UtBotEval` application is the core of our system. Its structure is described with the UML class diagram in Figure 3.3. 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

1. www.epicgames.com/unrealtournament/

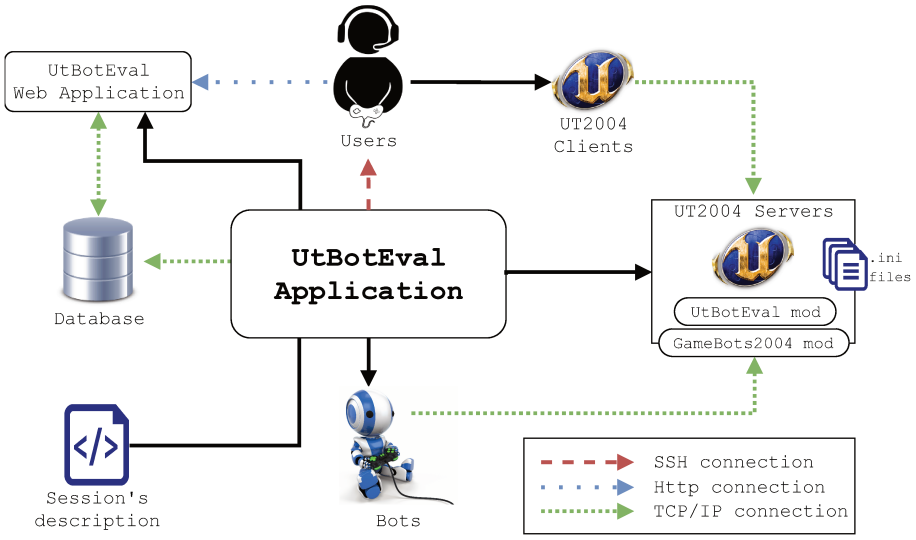


Figure 3.2 – System architecture of the UtBotEval system

(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

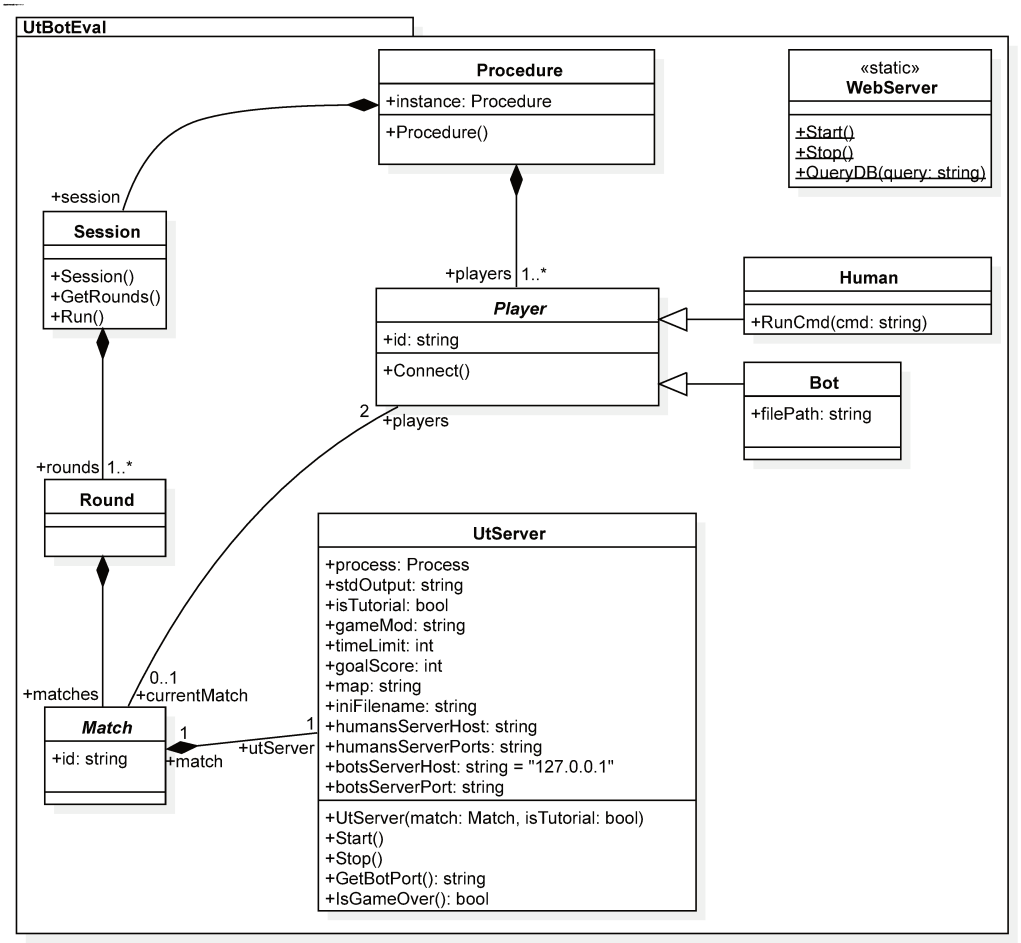


Figure 3.3 – UML class diagram of the UtBotEval Framework

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 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.

3.2.2 UtBotEval Mod for UT2004

Unreal Tournament 2004 (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.

```
// EvalBotDeathMatch.uc
class EvalBotDeathMatch extends BotDeathMatch;

var string Characters[32];
var int numCharacters;
var array<string> FemaleCharacters;
var array<string> MaleCharacters;

function string getCharacter(bool bIsFemale)
{
  local int i,index;
  local bool found;
  local string newCharacter;

  TryAnotherCharacter:
  found = false;
  if(bIsFemale)
  {
      index = Rand(default.FemaleCharacters.Length - 1);
      newCharacter = default.FemaleCharacters[index];
  }
  else
  {
      index = Rand(default.MaleCharacters.Length - 1);
      newCharacter = default.MaleCharacters[index];
  }
  for(i=0;i<numCharacters;i++)
  {
    if(Characters[i]~= newCharacter) goto TryAnotherCharacter;
  }
  Characters[numCharacters+1] = newCharacter;
  ++numCharacters;
  return newCharacter;
}

function string ChangeCharacter(string Options, string Character)
{
  local string Pair, Key, Value, Result;
  Result = "";

```

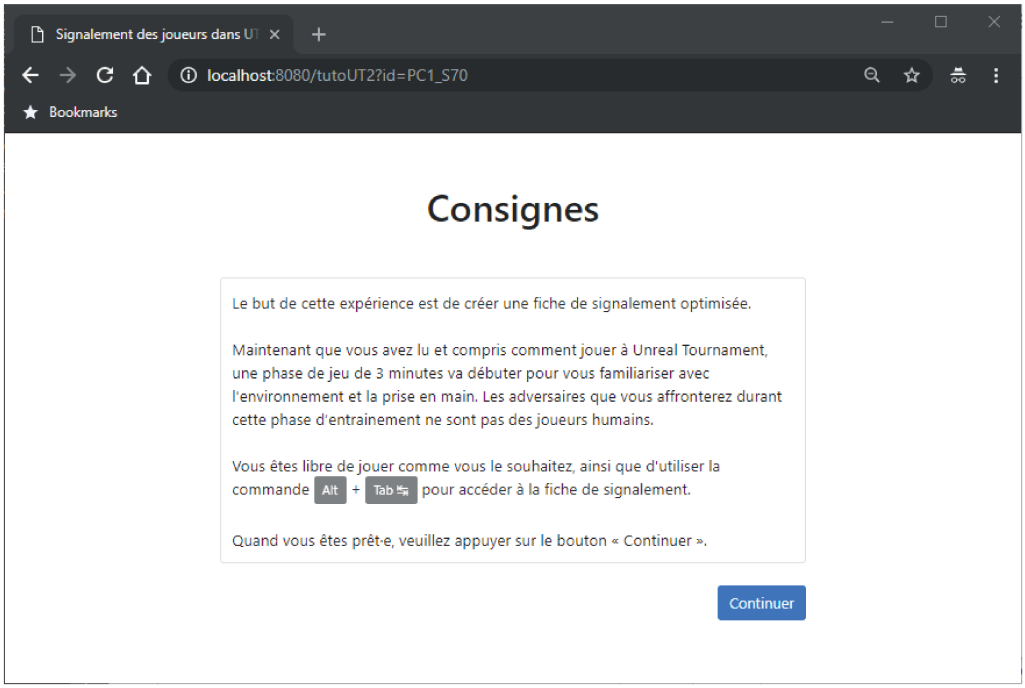
```
41    J0x08:
42    if(GrabOption(Options, Pair))
43    {
44      GetKeyValue(Pair, Key, Value);
45      if(Key ~= "Character")
46        Result = (Result $ "?Character=") $ Character;
47      else
48        Result = (((Result $ "?") $ Key) $ "=") $ Value;
49      goto J0x08;
50    }
51    return Result;
52  }
53
54  //Overriding the main function for adding bot to the game
55  function RemoteBot AddRemoteBot ([...])
56  {
57    local RemoteBot NewBot;
58    local string Character;
59
60    [...] //original GameBots code
61    // Get a random sex
62    Newbot.PlayerReplicationInfo.bIsFemale = bool(Rand(2));
63    // Change the skin of the bot
64    Character = getCharacter(NewBot.PlayerReplicationInfo.bIsFemale);
65    if (Character != "")
66      NewBot.DesiredSkin = Character;
67    else
68      NewBot.DesiredSkin = "ThunderCrash.JakobM";
69    // Change the name of the bot
70    ChangeName(NewBot, Character, true);
71    [...]
72    return NewBot;
73  }
74
75  //Overriding the function for spawning an epic bot
76  function bool AddEpicBot([...])
77  {
78    local Bot NewBot;
79    local string Character;
80
81    [...]
```

```
82    NewBot.PlayerReplicationInfo.bIsFemale = bool(Rand(2));
83    if ( NewBot != None )
84    {
85      Character = getCharacter(NewBot.PlayerReplicationInfo.bIsFemale);
86      NewBot.SetPawnClass("", Character);
87      [...]
88    }
89    ChangeName(NewBot, Character, false);
90    [...]
91    return true;
92  }
93
94  //Overriding the function called when new human player enters the game
95  event PlayerController Login( [...] )
96  {
97    local PlayerController Loging;
98    local GBReplicationInfo repInfo;
99    local string Character;
100
101   Loging = super.Login( Portal, Options, Error );
102   Loging.PlayerReplicationInfo.bIsFemale = bool(Rand(2));
103   Character = getCharacter(Loging.PlayerReplicationInfo.bIsFemale);
104   Options = ChangeCharacter(Options, Character);
105   repInfo = class'GBReplicationInfo'.Static
106             .SpawnFor(Loging.PlayerReplicationInfo);
107   repInfo.MyPRI = Loging.PlayerReplicationInfo;
108   ChangeName(Loging, Character, true);
109   RemoteNotifyLoging(Loging);
110   return Loging;
111 }
112
113 defaultproperties
114 {
115   MaleCharacters(0)="Outlaw"
116   MaleCharacters(1)="Kane"
117   [...]
118   FemaleCharacters(0)="Zarina"
119   [...]
120 }
```

3.2.3 UtBotEval Web Application

The web application is a simple collection of web pages connected to a database. These web pages allow to present the instructions of the experiment (see Figure 3.4), the tutorial of the video games, and the questionnaires. The answers to the questionnaires are collected and saved in the database.

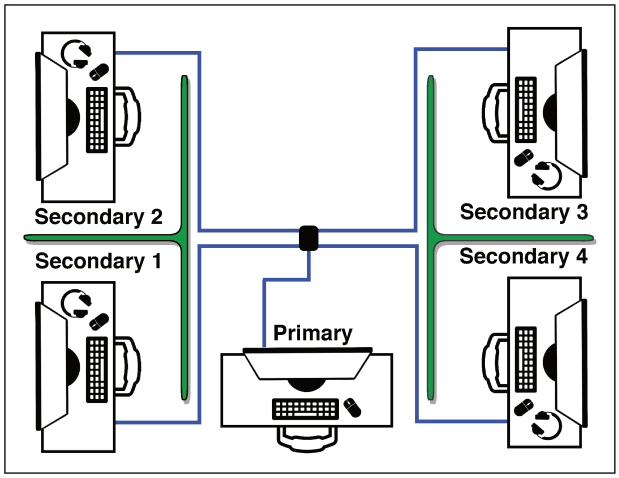
Figure 3.4 – Screenshot of one of the web application pages viewed by a participant



3.2.4 Physical Arrangement

Our new design greatly simplifies the physical arrangement of the competition. In its original version, the BotPrize required two separate rooms. In the first room there was one computer for each server, and one for each confederate while in the second room there was one computer for each judge. Our design requires only five computers (see Figure 3.5), all located in one room and connected to a Local Area Network (LAN). The primary computer is reserved for the investigator and the four secondary ones are intended for the participants. Secondary machines are

Figure 3.5 – Physical arrangement for the experiment



placed as far as possible from each other and headphones are provided to prevent the participants from hearing the other players and the sounds of the keyboard keys and mice that could give an indication of the fights in progress. Room dividers are placed to isolate the participants and to stop them from seeing the others' progression in the experiment. These precautions allow us to make sure that players do not guess when they play against another player in the room. The computers are equipped as follows:

The primary computer:
- The UtBotEval application.
- UT2004 and the UtBotEval mod.
- The executable files for the bots.
- The UtBotEval web application.
- A web server.

The secondary computers:
- UT2004.
- An SSH server.
- A web browser.

3.3 Experiment Methodology

We conducted this experiment to verify if our hypothesis is valid. As a reminder, 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.

3.3.1 Participants

Four groups of four students (16 participants) from the Brest National School of Engineering [2] participated in the experiment. The participants were all volunteers and no compensation was provided for their engagement.

3.3.2 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 [3]:

> This experiment lasts approximately one 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.

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 is given in Table 3.1 and was generated partially randomly to meet this constraint.

3.4 Results

To evaluate the experiment, we analysed the answers given in the final questionnaire. Unfortunately the results were not as expected. Out of six-

2. École nationale d'ingénieurs de Brest (ENIB): www.enib.fr
3. Translated from French

Table 3.1 – Passing order of participants

		Rounds															
Sessions	Participants	1		2		3		4		5		6		7		8	
1	h0	h1	A	h3	B	b1	C	b2	D	h2	E	b3	F	b4	G	b0	H
	h1	h0	A	b3	G	b4	H	b2	F	b0	B	h2	C	h3	D	b1	E
	h2	b2	G	b4	B	b1	F	b0	D	h0	E	h1	C	b3	A	h3	H
	h3	b1	A	h0	B	b3	G	b0	F	b2	E	b4	C	h1	D	h2	H
2	h0	b2	H	b1	D	b4	F	h1	C	b3	E	h3	A	h2	B	b0	G
	h1	b2	D	b3	B	b4	E	h0	C	b0	A	h2	F	h3	H	b1	G
	h2	b4	A	b0	G	b3	E	b1	H	b2	C	h1	F	h0	B	h3	D
	h3	b0	C	b1	B	b2	E	b3	F	b4	G	h0	A	h1	H	h2	D
3	h0	b4	D	b1	A	h1	F	h2	H	h3	G	b0	E	b3	C	b2	B
	h1	b4	B	h2	G	h0	F	b2	H	b3	C	b0	D	h3	A	b1	E
	h2	h3	E	h1	G	b2	C	h0	H	b3	A	b1	B	b4	F	b0	D
	h3	h2	E	b4	C	b0	B	b1	F	h0	G	b3	H	h1	A	b2	D
4	h0	h2	G	h1	E	b0	H	b3	D	b2	A	b1	C	b4	B	h3	F
	h1	b1	D	h0	E	h2	H	b2	G	h3	A	b4	F	b0	B	b3	C
	h2	h0	G	b4	A	h1	H	h3	B	b1	E	b3	F	b2	D	b0	C
	h3	b1	C	b2	E	b0	D	h2	B	h1	A	b3	G	b4	H	h0	F

Map names :

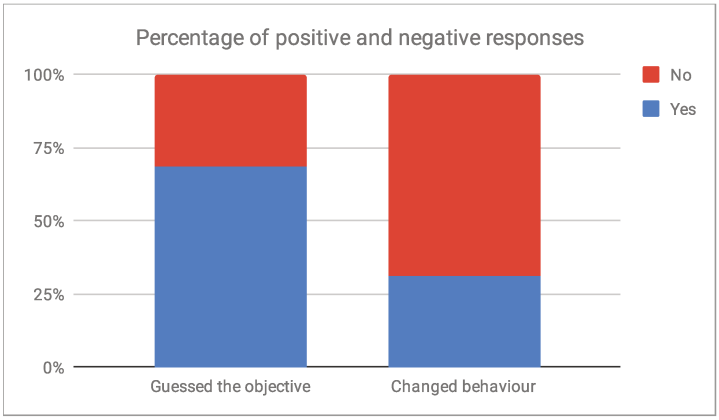
A : DM-1on1-Spirit
B : DM-1on1-Idoma
C : DM-Insidious
D : DM-1on1-Irondust
E : DM-Leviathan
F : DM-Gael
G : DM-1on1-Desolation
H : DM-1on1-Albatross

teen 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.

3.5 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 the Brest National School of Engineering are particularly familiar with video games as they are trained in computer pro-

Figure 3.6 – A 100% stacked bar chart with the responses of the final questionnaire



gramming 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 chapter.

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 chapters to follow.

3.6 Conclusion

In this chapter we presented our first protocol proposal to evaluate the believability of bots in the UT2004 video game. 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.

To build this new protocol we were inspired by the first version of the BotPrize competition to which we made some modifications. In addition to replacing the questionnaire, we integrated a training phase to allow participants to become familiar with the video game before starting the evaluation. We also changed the format of the matches by choosing to play only one-on-one games so as not to distract participants with a third player.

To facilitate the implementation of this evaluation protocol we have developed a system that partially automates the execution of the evaluation process. This facilitates the logistics of the evaluation and reduces the risk of mishandling by the investigator. In addition, the system is very flexible since the architecture is composed of 3 independent elements that can be modified as needed.

To check our approach we conducted an experiment where we executed our protocol and added an additional questionnaire at the end of it to check if the participants had guessed the purpose of the experiment and whether they had changed the way they played. The results obtained were not satisfactory but they led us to question the impact that the level of expertise of participants in video games can have on such an assessment method. We will therefore study this element in the next chapter.

CHAPTER 4

4. INFLUENCE OF THE JUDGES' EXPERTISE

Part of this chapter was published in the 17th Cyberworlds Conference proceedings (Even et al., 2018)

CW2018

In the previous chapter, our attempt to mask to the judges what was actually evaluated failed and we wondered if this was due to the participants' familiarity with video games. It therefore seems interesting to investigate this element and more specifically to study the impact that the level of expertise of judges in video games could have on their ability to distinguish bots from human players. In this chapter, we do not intend to try to avoid gameplay modifications. Therefore, we simplified the questionnaire of our previous protocol to keep only the question used to judge the opponent. 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.

At an opportune time, we had the possibility to join the competition organisation committee of the French Association of Artificial Intelligence (AFIA [1]). Every summer, this committee is in charge of organising a com-

1. https://afia.asso.fr/

petition at a national conference coordinated by the association. The competition features a different topic each year but the main themes alternate from one year to another and are: the application of artificial intelligence in 1) robotics and 2) video games. We participated in the organisation of the 2017 edition, called the BotContest [2], which took place at the artificial intelligence platform in Caen (PFIA'17 [3]). We took advantage of this event to organise a competition, where 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 implement our protocol and obtain useful information regarding the judges and their level of expertise in video games. In the rest of the chapter, the term "participants" refers to individuals who participated in the jury and not the competitors as they were not present during the competition.

First, in section 4.1, we will present the changes we made to our evaluation protocol and the implementation of our computer system to collect data on the game and its players. Then we will present in detail in section 4.2, the metodology of the experiment that we realized during the final of the BotContest competition. The results of our analysis are given in section 4.3. Finally, we give a discussion of the results in section 4.4 and we conclude in section 4.5.

4.1 Model Modifications

Match ending condition: Our first choice for the competition was to set a "TimeLimit" (the maximum duration of the game) to ensure that all the participants would play the same amount of time with all their opponents. However, we observed that the number of times the participants met their opponent varied significantly from one match to another. Similarly, when setting a "GoalScore" (the score required to win the match) to n, the number of times the players would meet each other could vary from n (i.e. one player got all the points) to $(2n-1)$ (ie. the game is tied until the last shot). In order for the competition to be fairer, we changed the behaviour of the GoalScore. We decided to 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 with the

2. http://afia-competitions.fr/botcontest/
3. https://pfia2017.greyc.fr/

GoalScore parameter, the game ends automatically. We do not count "suicides" among the frags since they are generally not due to the opponent. Suicides can occur either by falling into a hole, lava or acid, by shooting yourself or getting hit by your own weapons blast. We also keep a TimeLimit as a security to make sure the game does not last too long for logistical reasons. In order not to confuse the game's original GoalScore and our modified version we will call this parameter "FragLimit" in the rest of the chapter.

Assessment method: For this proposal we simplified the questionnaire by keeping only the question used to judge the opponent. We therefore ask participants to:

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

4.1.1 Implementation

Since the evaluation is subjective in nature, it is important to collect as many judgements as possible. In order to save time and to have a maximum of participants, we reused the framework presented in chapter 3. The only modifications that we had to bring to the system was the update of the questionnaire in the web application and the implementation of the new ending condition in our UT2004 mod (see code below). We also used our mod to log some information regarding the match such as its duration and the scores of the players.

```
1  // EvalBotDeathMatch.uc
2  class EvalBotDeathMatch extends BotDeathMatch;
3
4  var FileLog EvLog;
5  var string EvLogFileName;
6  var int fragCount;
7
8  // Function automatically called after the beginning of the game
9  // Overridden to create the new log file
10 function PostBeginPlay()
11 {
```

```
12    [...]
13    EvLogFileName = ((((((((((("BotEval" $ string(Level.Day)) $ "/")
14      $ string(Level.Month)) $ "/") $ string(Level.Year)) $ ":")
15      $ string(Level.Hour)) $ ".") $ string(Level.Minute)) $ ".")
16      $ string(Level.Second);
17    StartEvLog();
18  }
19
20  // Function automatically called after the beginning of the game
21  // Overridden to close the log file by calling EndEvLog()
22  function EndGame(PlayerReplicationInfo Winner, string Reason )
23  {
24    log("In EndGame "$Winner.PlayerName$" won the match.");
25    if(EvLog != none)
26      EndEvLog(Winner,Reason);
27    Super.EndGame(Winner,Reason);
28  }
29
30  // Main function for adding bot to the game
31  // Overridden to fill in the anonymisation tag
32  function RemoteBot AddRemoteBot([...])
33  {
34    [...]
35    // change the bot's name and write it in the log file :
36    WEvLog("    <player>");
37    WEvLog("      <type>bot</type>");
38    WEvLog("      <originalName>" $ clientName $ "</originalName>");
39    changeName( newBot, Character, true );
40    WEvLog("      <newName>" $ NewBot.PlayerReplicationInfo.PlayerName
41                                $ "</newName>");
42    WEvLog("    </player>");
43    [...]
44  }
45
46  //[...] Same modifications for the AddEpicBot() and Login() functions
47
48  // This function creates a new log and initialises it
49  function StartEvLog()
50  {
51    if(EvLog != none)
```

```
52   {
53     EvLog.CloseLog();
54     EvLog.Destroy();
55     EvLog = none;
56   }
57   EvLog = Spawn(class'FileLog');
58   WEvLog("<?xml version=\"1.0\" encoding=\"UTF-8\"?>");
59   WEvLog("<game id=\"" $ DemoCommand $"\">");
60   WEvLog("  <anonymization>");
61 }
62
63 // Closes the anonymization tag and logs match ending reason and scores
64 function EndEvLog(PlayerReplicationInfo Winner, string Reason)
65 {
66   local Controller P;
67   if(Winner!=None){
68     for ( P=Level.ControllerList; P!=None; P=P.nextController )
69     {
70       if(!PlayerController(P).IsSpectating()){
71         if(P.PlayerReplicationInfo.Score > Winner.Score)
72           Winner = P.PlayerReplicationInfo;
73       }
74     }
75   }
76   WEvLog("  </anonymization>");
77   if(Winner!=None){
78     WEvLog("  <endgame reason=\""$Reason$
79                    "\" winner=\""$Winner.PlayerName$
80                    "\" game_duration=\""$ElapsedTime$"\">");
81   }else{
82     WEvLog("  <endgame reason=\""$Reason$
83                    "\" winner=\"None\" game_duration=\""$ElapsedTime$"\">");
84   }
85   for ( P=Level.ControllerList; P!=None; P=P.nextController )
86   {
87     if(!PlayerController(P).IsSpectating()){
88       PlayerReplicationInfo PRI = P.PlayerReplicationInfo;
89       WEvLog("    <player name=\""$PRI.PlayerName$"\">");
90       WEvLog("      <oldname>"$PRI.OldName$"</oldname>");
91       WEvLog("      <score>"$PRI.Score$"</score>");
```

```
92        WEvLog("       <kills>"$PRI.Kills$"</kills>");
93        WEvLog("       <deaths>"$PRI.Deaths$"</deaths>");
94        WEvLog("       <numLives>"$PRI.NumLives$"</numLives>");
95        WEvLog("    </player>");
96      }
97    }
98    WEvLog("  </endgame>");
99    WEvLog("</game>");
100   EvLog.CloseLog();
101   EvLog.Destroy();
102   EvLog = none;
103 }
104
105 // Write the text (txt) passed as an argument in the log file.
106 function WEvLog(string txt)
107 {
108   if(EvLog == none) StartEvLog();
109   else
110   {
111     EvLog.OpenLog(EvLogFileName);
112     EvLog.Logf(txt);
113     EvLog.CloseLog();
114   }
115 }
116
117 // Update the fragCount after each kill
118 function ScoreKill(Controller Killer, Controller Other)
119 {
120   if(Killer != Other && Killer != None) fragCount++;
121   super.ScoreKill(Killer,Other);
122 }
123
124
125 // Function automatically called to check if the score means the game ends
126 // Overridden to use our new ending condition :
127 // GoalScore = FragLimit : number of frags required to stop the match.
128 function CheckScore(PlayerReplicationInfo Scorer)
129 {
130   local controller C;
131   if ( Scorer != None )
```

```
132   {
133     if ( (GoalScore > 0) && (fragCount >= GoalScore) ){
134       for ( C=Level.ControllerList; C!=None; C=C.NextController ){
135         if ( (C.PlayerReplicationInfo != None)
136              && (C.PlayerReplicationInfo != Scorer)
137              && (C.PlayerReplicationInfo.Score > Scorer.Score) )
138           EndGame(C.PlayerReplicationInfo,"FragLimit");
139         else EndGame(Scorer,"FragLimit");
140       }
141     }
142   }
143 }
```

4.2 Experiment Methodology

Our objective was to use this competition to analyse the judges' expertise according to their video game habits. In this section we detail the characteristics of this experiment.

4.2.1 Participants

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

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.

4.2.2 Procedure

On the participants' arrival, a web page was already opened with the following indications (translated from French):

> "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. The second phase of the experiment consisted of four stages where the participants played a match of UT2004 with the BotContest mod and then filled the judging form after each match. During the four rounds, the participants would face the three bots and one of the other participants. Obviously, this information was kept secret and participants only knew that they would face a random number of bots and humans, in a random order. In the final phase, participants were invited to complete a questionnaire that collected personal information regarding their gaming habits (see subsection 4.2.4 for a detailed description.).

4.2.3 Independent Variables

The maps: For this experiment, we used four different maps from the game: DM-1on1-Albatross, DM-1on1-Spirit, DM-1on1-Idoma and DM-Gael. We selected these maps for their small size as it is the most appropriate for one-on-one deathmatch games. For instance, the DM-Gael map (see Figure 4.1) was chosen for its particularity of having only one main room with a fairly large and deep pit in the middle. Floating in the centre of this pit is a platform where power-ups can spawn. Reaching this pickup comes at a risk, as falling down the pit will result in death.

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 during the qualification process that some bots could not maintain a believable behaviour over the long term. Past three minutes, some began to have repetitive and predictive behaviour such as repeated back and forth or using always the same route or the same attack strategy. For these reasons we decided that the duration of a match should be greater than three minutes.

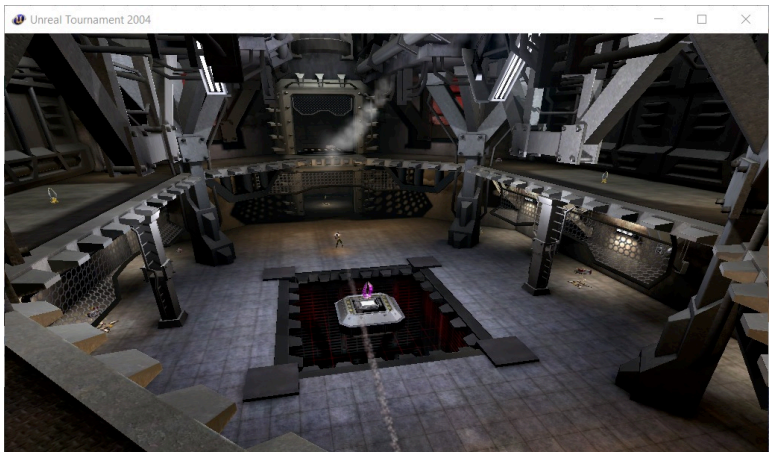


Figure 4.1 – Screenshot of the DM-Gael map from UT2004

The FragLimit: It was set to 10 after extensive testing. We previously used a FragLimit of 5 and noticed that the matches lasted 2:30 minutes on average. Therefore, we decided to double the FragLimit to obtain a match duration closer to 5 minutes on average.

4.2.4 Measures

Thanks to our framework, we were able to automatically record match information in a database so that we could easily process it with queries. For each game we collect the following data: 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 two scores: 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 had to fill a four-questions questionnaire to evaluates their video game expertise [4]:

4. Translated from French, original version in Appendix D

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 (e.g. games on CDs or on-line)
   - Console (e.g. Xbox, Playstation or wii)
   - Hand-held game console (e.g. Game Boy, PSP)
   - Arcade game (e.g. coin-operated entertainment machine installed in public businesses)
   - Other device (e.g. Mobile phone or MP3 player)
3. What types of games do you play?
   A) First-Person Shooter (e.g. counter-strike)
   B) Strategy games (e.g. Age of empire)
   C) Platform games (e.g. Rayman)
   D) Adventure, Action Games (e.g. Assassin's Creed)
   E) RPG: Role Playing Game (e.g. Final Fantasy)
   F) Educational games (e.g. Adibou)
   G) Management Games (e.g. Zoo Tycoon)
   H) Simulation games (e.g. Sims)
   I) Sports Games (e.g. Fifa, PES)
   J) Racing Games (e.g. Grand Turismo, Mario Kart)
   K) MMORPG = Massively multi-player on-line role-playing game (e.g. World of Warcraft)
   L) Physical or sports games (e.g. Wii, Kinect, Playstation Move)
4. Do you play :
   - Alone
   - With virtual 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.

4.2.5 Statistical Techniques Used

Thanks to the data collected in our database we were then able to carry out some analyses. Here are the statistical methods we used:

Kruskal-Wallis Test: 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).

Contingency Table: Also known as a cross tabulation or crosstab, it is a type of table that displays the frequency distribution of two categorical variables

Chi-Square Test This method is applied to examine whether rows and columns of a contingency table are statistically significantly associated.

Correspondence Analysis (CA) This method provides a graphic method of exploring the relationship between row and column variables in a contingency table. It is an extension of Principal Component Analysis (PCA) suited to handle qualitative variables (or categorical data).

Multiple Correspondence Analysis (MCA) It is an extension of the CA which allows to analyse the pattern of relationships of several categorical dependent variables.

4.3 Results

4.3.1 Competition Results

The results of the competition are given in Table 4.1. We note that the scores for the bots are all negative which means that none of them passed the test. The obtained ranking is the same whether we take into account the certainty scale or not. Given that we used this scale for the simple purpose of discouraging participants from not doing the cognitive work (as we explained in chapter 3) and that analysing data at the same time with and without the degree of certainty gave us the same results, we decided to present in this chapter only the results using the humanness score without the degree of certainty for the sake of simplification and

to avoid repetitions. To analyse the difference between the humanness score for humans and bots, we performed a T-Test which gave us a p-value of 9.3e-7 indicating that the difference is significant.

Table 4.1 – 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

From the data we collected during the competition, we were able to study some interesting features of the protocol. First of all, the bar plot in Figure 4.2 shows the repartition of the matches duration for each map. The duration of matches was discretised into five classes: matches lasting less than 2; 3; 4; and 5 minutes; and games ending with the time-limit condition which was set at 5 minutes. We note that the duration of the match differs from one map to another. To validate 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 4.3). 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 4.4 : the shorter the matches, the lower the score. However, 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 following p-values: (a) 0.67, (b) 0.52, (c) 0.76. This does not allow us to reject the

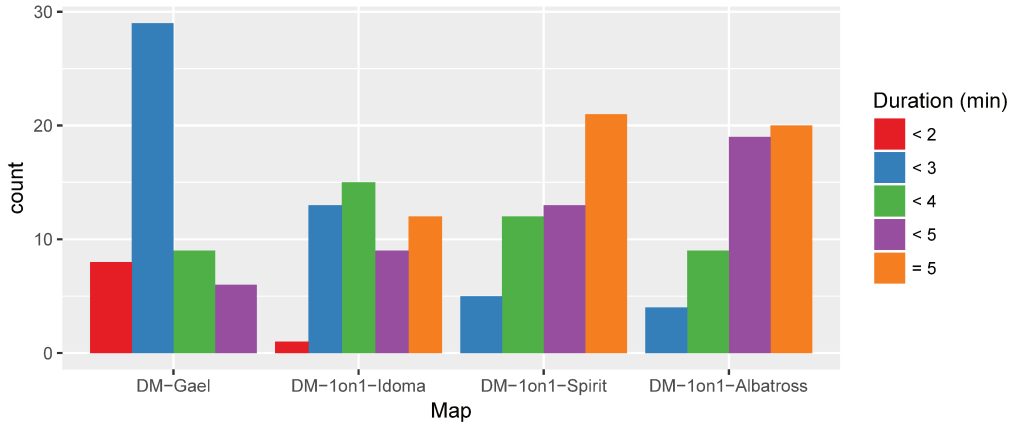


Figure 4.2 – Bar plot of the match durations (in minutes) depending on the maps

null hypotheses and leads us to conclude that there is no link between these elements and the humanness score.

4.3.2 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 while the worst were wrong at least 3 times out of 4.

4.3.2.1 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 (see Figure 4.5) is rather interesting since it shows that the best judges tend to play everyday, the worst tend to never play, and intermediate judges play occasionally.

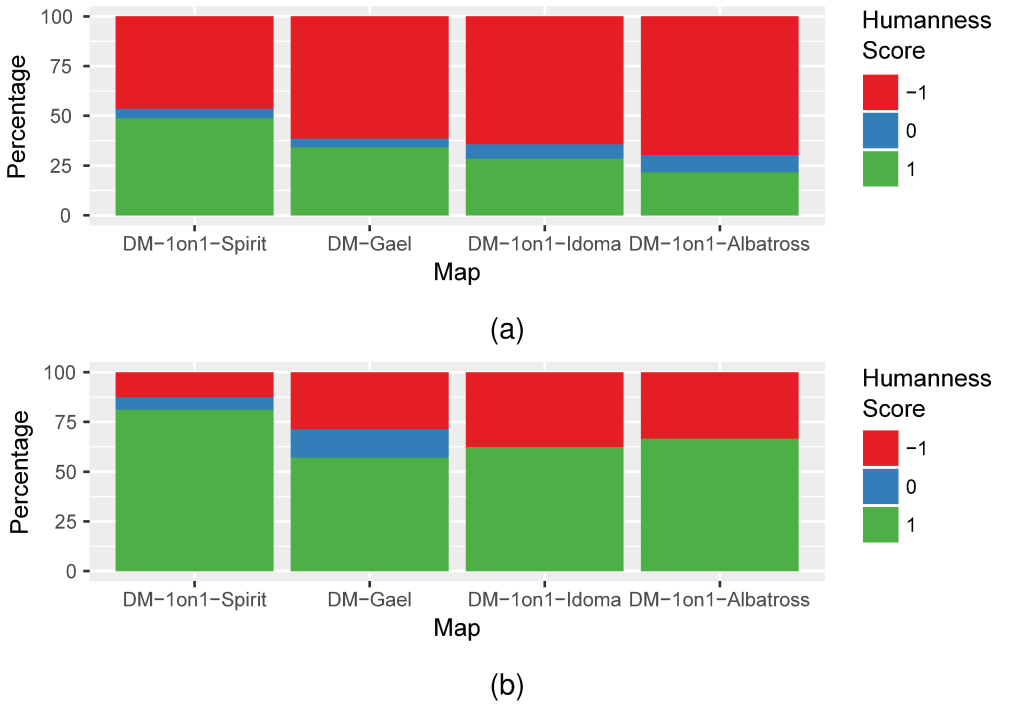


Figure 4.3 – Bar plot of the humanness score for (a) bots and (b) humans depending on the maps.

4.3.2.2 Usual Type of Game

Using the same method as above we obtained a chi square of independence between the two variables of 31.60 (p-value = 0.024). We can therefore reject the null hypothesis and 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 4.6) 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 4.2.4. 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

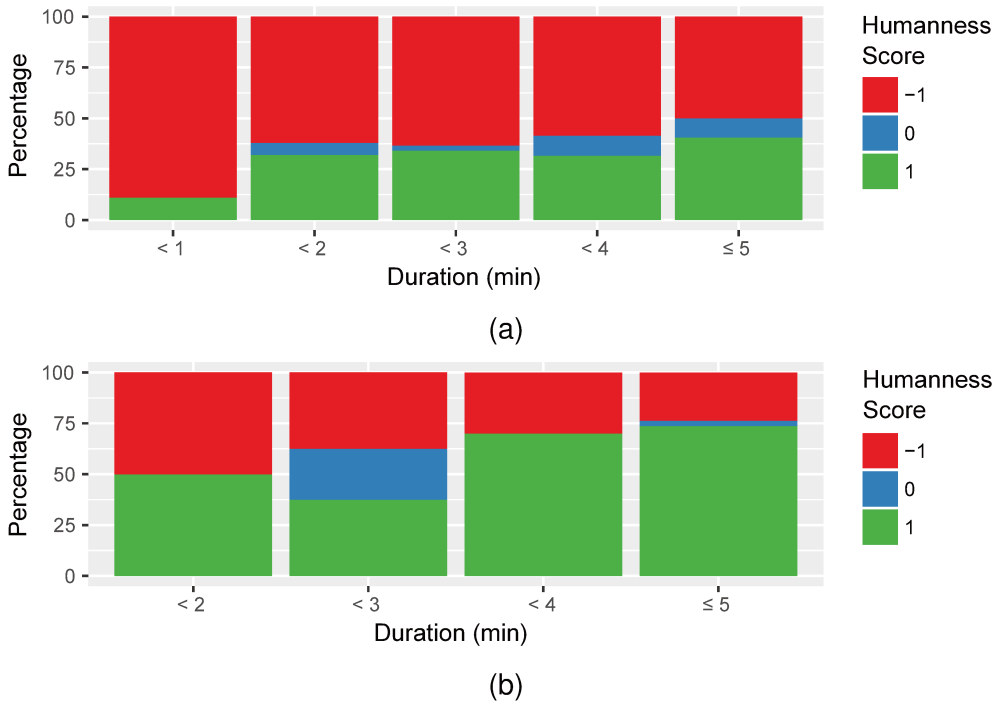


Figure 4.4 – Bar plot of the humanness score for (a) bots and (b) humans depending depending on the match durations (in minutes).

contain shooting phases, or at least, this is very rare.

4.3.2.3 Usual Devices

The distribution of the answers chosen by the participants concerning the devices used is similar for all levels of expertise (see Table 4.2). There is therefore no link between these two elements.

Table 4.2 – 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

4.3.2.4 Usual player types faced

Table 4.3 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 4.7). 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 4.3.

Table 4.3 – 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 [5]
Best	100	60	70	70	90
Intermediate	80	33	28	48	48
Worst	70	40	40	70	40

4.4 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

5. Physically Present Players

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 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 takes training to acquire the necessary skills to be able to master this type of

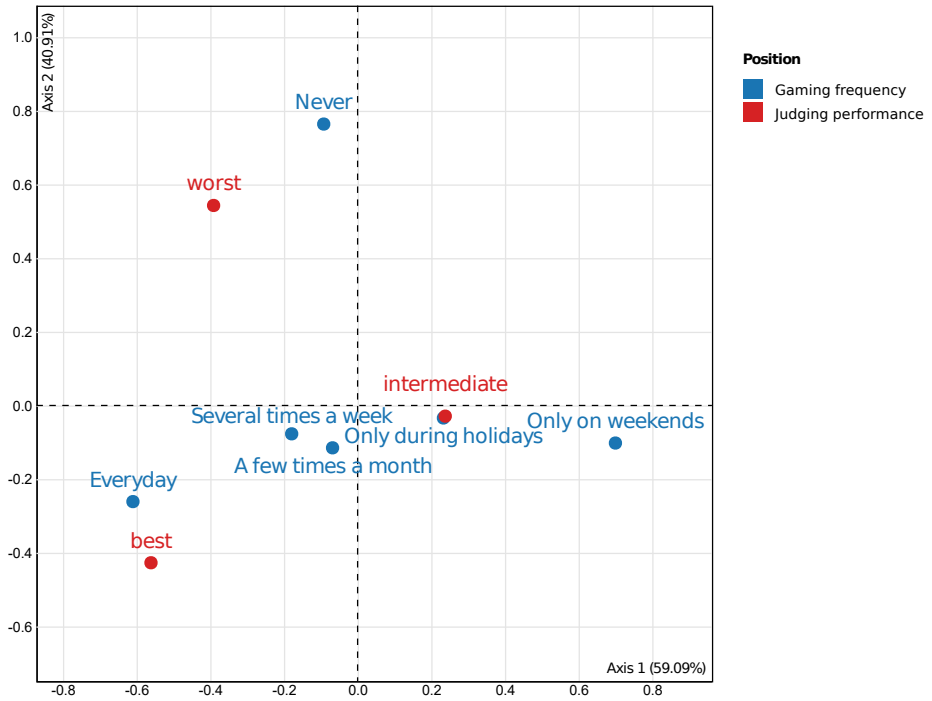


Figure 4.5 – Correspondence analysis factor map

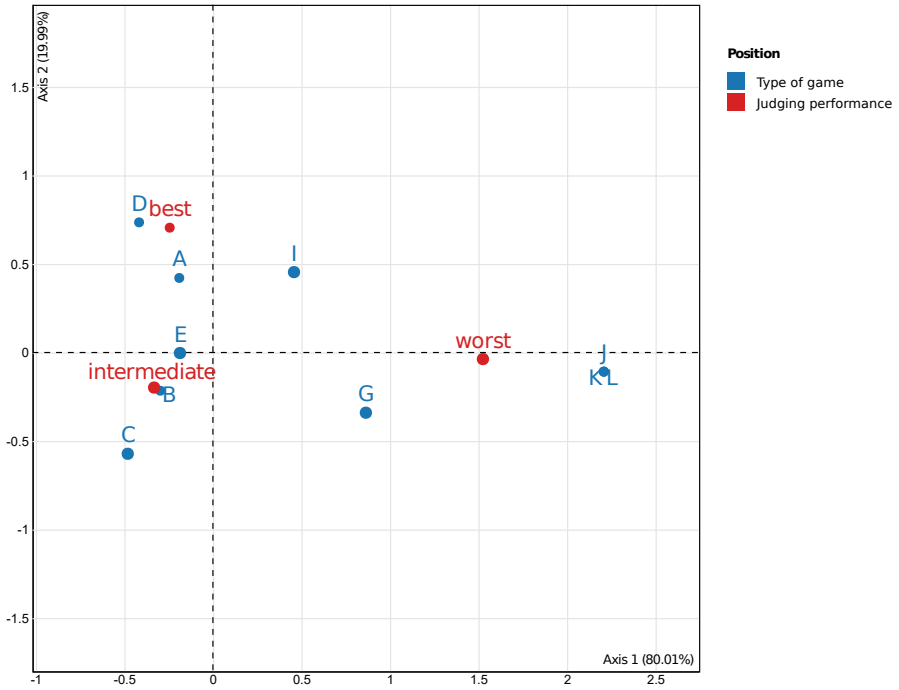


Figure 4.6 – Correspondence analysis factor map

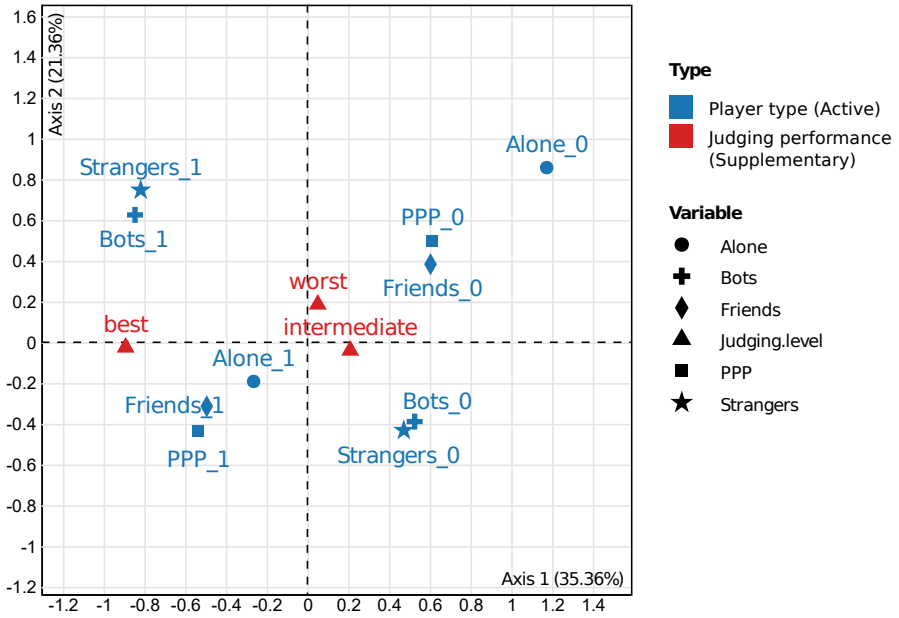


Figure 4.7 – Multiple correspondence analysis plot for dimensions 1 and 2

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.

4.5 Conclusion

In this chapter, we presented the modifications we made to our evaluation protocol and to our computer tool. As you may have noticed, only one of the elements of our tool has been modified and it has had no impact on the rest of our system which illustrates its flexibility. Thanks to our system, we were able 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 2008 BotPrize competition (Philip Hingston, 2009). We took advantage of this event to collect and analyse data concerning both certain characteristics of the protocol as well as the gaming habits of the volunteers who participated in the jury of the 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.

We also saw that participants with the most difficulty distinguishing bots from humans are novice players. We believe it is important that the jury of such a competition be composed of players of different levels. However, giving judges the opportunity to give their opinion only when they wish could be a better approach. Indeed, for some novice players, simply learning to play a new game can be quite overwhelming and asking them to do an extra task may seem too difficult.

Finally, we observed that some judges did put strategies in place to unmask the nature of their opponents rather than play. Conducting the

evaluation in the form of a competition may worsen these behaviours as volunteers being invited to join the jury feel unconsciously pushed to judge rather than play. We continue to think that it is important for judges to ignore the purpose of the experiment.

CHAPTER 5

5. REPORTING SUSPECTED CHEATERS

Until now, the believability of bots was directly evaluated using different methods, which did not allow to obtain convincing results. One of the main problem being that the gameplay can be modified by these so-called "first-person" assessment methods, as we have seen in the previous chapters. Players are more focused on judging than playing the game, which introduces new behaviours in the game. In this chapter, 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; D. T. Campbell and Fiske, 1959). Current methods of assesing 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. (Widaman, 2018). 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 re-

port prejudicial behaviours faced when playing a game. Most of the time, these systems offer many options to report abuses (see Figure 5.1), 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 (AFK). 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 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 (as in Figure 5.2) 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. This new model is presented in section 5.1. To validate our hypothesis, we carried out an experiment whose methodology is presented in section section 5.2. To achieve this experiment, we used once again, the computer system presented previously. The section 5.3 presents the results we obtained which are then discussed in section 5.4. Finally, we conclude this chapter in section 5.5.

5.1 Model

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. Indeed, this feature is relatively recent and did not exist at the time of this game's release. We therefore developed a reporting system for this game by taking inspiration from existing ones in other video games.

Figure 5.1 – Reporting form from the video game Fortnite

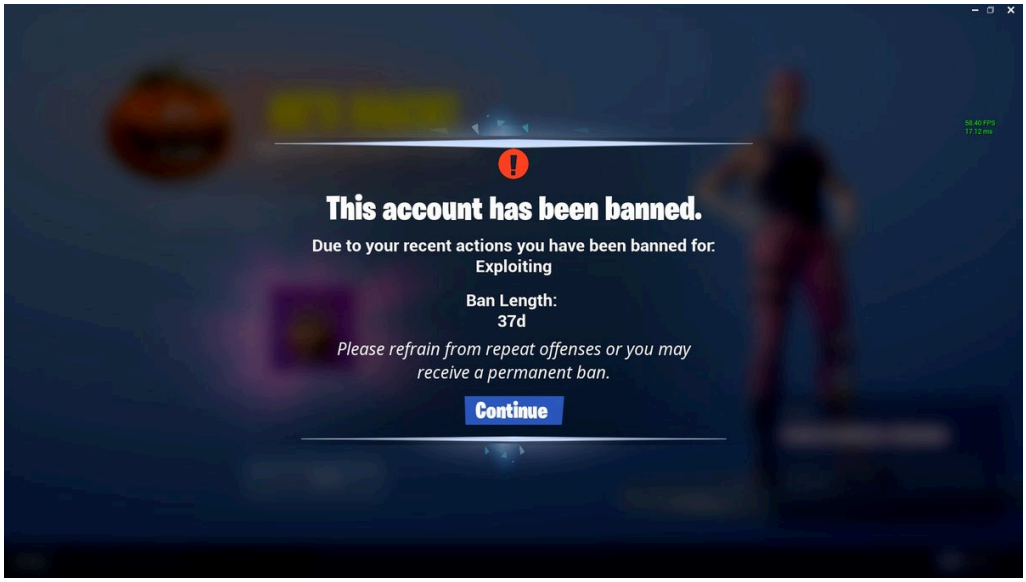


Figure 5.2 – Ban notification from the video game Fortnite

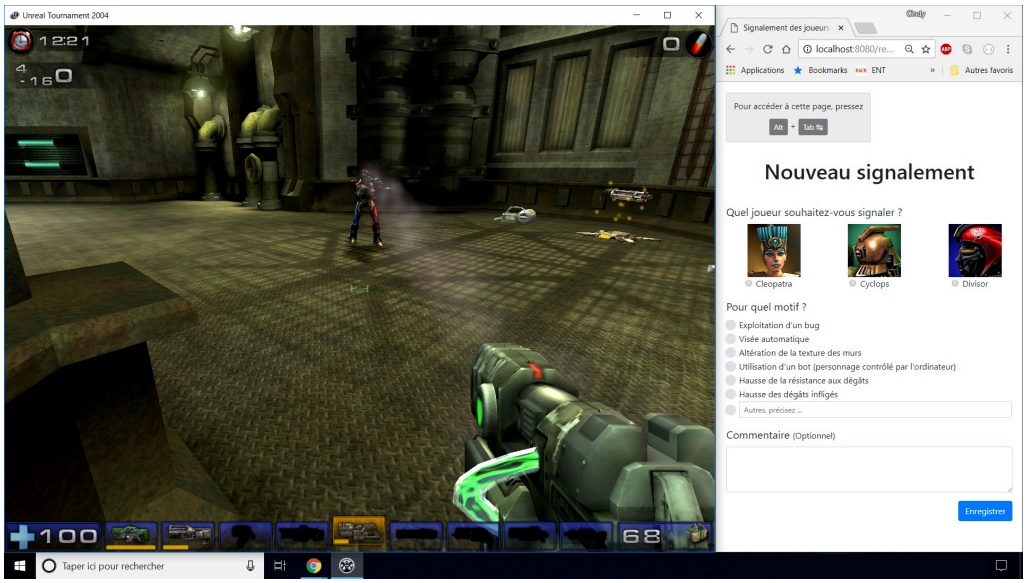


Figure 5.3 – Screenshot illustrating the position of the windows: on the left the video game UT2004, on the right the reporting form opened in a browser (see section E.2 for a larger image)

First of all, we were interested in the various solutions that 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. Regarding the first solution, we found that it was not suitable for a game such as an 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 (see Figure 5.3). To access it, the player simply has to change the active window with the Alt + ⇤⇥ keys combination.

Then, we looked for what options we would integrate into the reporting form. To determine them, we studied the codes of conduct of sev-

eral 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 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.

5.2 Experiment Methodology

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. We describe the experience in detail in the rest of this section.

5.2.1 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.

5.2.2 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 (similar to the one described in 4.2.4; see appendix E.1). 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 five 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.

5.3 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 5.4). The bivariate Wilcoxon test gave a p-value of 0.0547 which does not allow us to reject the null hypothesis. 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.

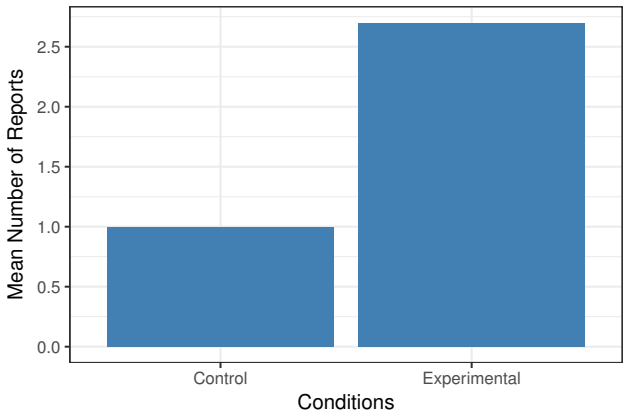


Figure 5.4 – Mean number of reports depending on the condition

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-value = 0.2021).

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-value = 0.03362). 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.

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”. 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 5.5, the two groups do not seem to be significantly different, this was confirmed by a Wilcoxon test which gave a p-value of 0.3119.

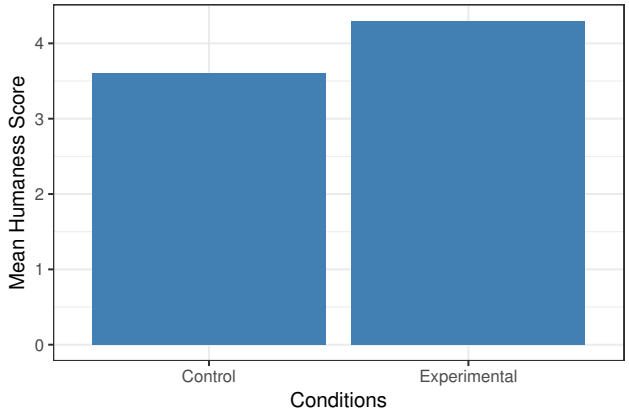


Figure 5.5 – Mean humanness score for bots depending on the condition

Participants were asked to indicate how many human players they thought they faced. They could choose a value between 0 and 3. Their answers do not seem to be significantly different (p-value = 0.9544) between the control group (2.57 ± 0.53) and the experimental group (2.6 ± 0.52). The same question was asked regarding the number of bots. Again, the difference is not significant (p-value = 0.8687) between the control group (0.71 ± 0.76) and the experimental group (0.6 ± 0.62).

They 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”). Their degree of certainty for the number of human player do not seem to be significantly different (p-value = 0.5469) 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-value = 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-value = 0.05766, cor = -0.7391) whereas for the experimental group (see Figure 5.6), a strong negative correlation seems to appear between the number of reports and the believability score (p-value = 0.0273, cor = -0.6898).

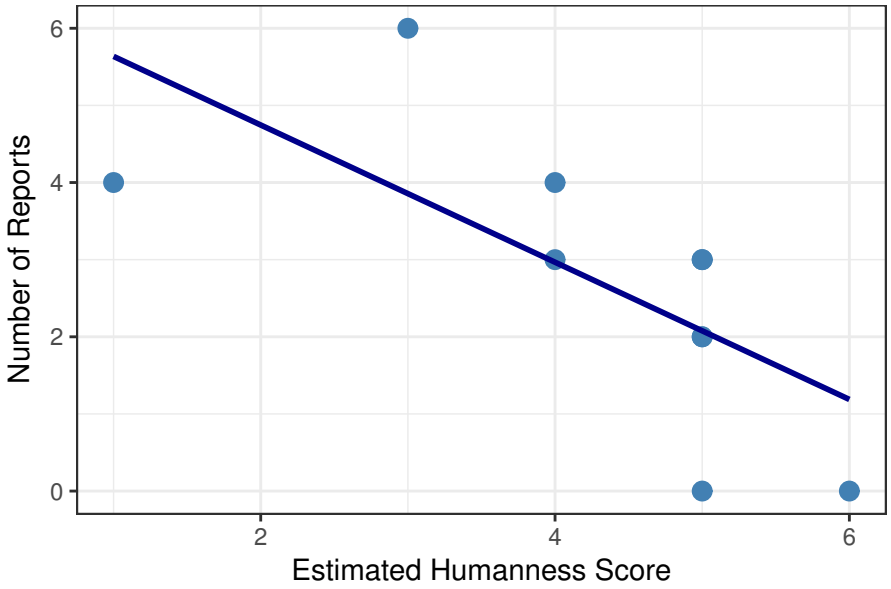


Figure 5.6 – Negative correlation between the bots' estimated humanness and the number of reports in the control group

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.

5.4 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 play-

ers 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 hypothesis 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 indi-

vidually. 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 chapter. 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.

5.5 Conclusion

Our desire to keep the nature of the game unchanged when evaluating the believability of the bots led us to propose a novel solution, based on the reporting systems frequently encountered in online multiplayer video games. In this chapter, we presented our model and its implementation for the video game Unreal Tournament 2004. Since this game does not have a reporting system by default, we have developed one in order to realise an experiment to validate our approach. Seventeen people participated in our experiment out of which ten were in the experimental condition (matches of 2 bots and 2 humans) and seven in the control condition (matches of 4 humans).

Despite the fact that the number of participants was quite small, we were able to make some interesting observations. In particular, it would appear that the participants in the experimental condition made more reports than those in the control condition. Also, there is a negative correlation between the number of reports and the believability score of the bots.

Thus, it would appear that it is possible to use this objective measure to evaluate the believability of the bots.

Some improvements are still possible. Regarding the evaluation method, it would be interesting to ask participants to evaluate the believability of all their opponents at the end of the experiment in order to study the existence of a direct link between reporting and believability. Then, regarding the reporting form, we focused on the particular case of FPS but it would be interesting to establish different lists of reporting reasons according to the types of video games.

Finally, we have noticed that with this new approach, the participants have not changed the way they play and have not put in place strategies to try to unmask the nature of their opponents. During our discussions with them after the experiment, we noticed that they did not realise that our goal was to evaluate the believability of the bots. This is very important for us since it guarantees a rigorous and unbiased evaluation.

CHAPTER 6

6. Conclusion and Future Work

6.1 Conclusion

The goal of this thesis was to put in place a rigorous protocol to evaluate the believability of virtual 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 virtual 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 dis-

perse 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 the PFIA17 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 fount 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.

6.2 Future Work

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. In particular, we used the video game Unreal Tournament 2004 throughout our study, however, it would be interesting to test our protocol with newer games. Our choice fell on this game because it was available to us and it was easy to integrate external bots to it. However, this game being too old, it does not include a reporting system by default. A new version of the game is currently in development (Unreal Tournament 4 [1]) and it should certainly have a reporting system as it is present in the majority of recent video games distributed by AAA publishers. This new version being very similar to the old one, we think that it will be possible to modify the game and to integrate bots as for UT2004. This would allow us to test our protocol with an updated version of the video game.

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.

Our method can be used not only by the scientific community but also by video game publishers. They could integrate different bots implementations on some of their game servers and use the player reports to determine which implementations are the most believable. From these tests results they could improve their bots until they are no longer reported. However, since video game publishers do not have much time to devote to such work, it would be interesting to set up test platforms. The idea would be to provide network text protocols for connecting to the game

1. https://www.epicgames.com/unrealtournament/

and control game characters as GameBots2004 (Bída et al., 2012) does for UT2004 and StarCraft II API[2] does for the game of the same name (Vinyals et al., 2017). This would allow researchers and independent developers to implement bots that could interact with the video game. Game publishers could then make a few servers available on these platforms so that everyone can connect his or her bots and where volunteer players could play. Any reports would then be sent directly to the bots which would allow comparisons and improvements of the different implementations.

We have noticed in our study that novice players often confuse expert players with bots and conversely, expert players confuse novices with bots. This phenomenon is common in online video games. For example, on one hand, novice players, called newbies or noobs, are often insulted by more experienced players who can not stand to lose because of the inexperience of their teammates. While on the other hand, expert players are accused of cheating because of their accuracy, speed or knowledge of the map. To avoid these conflicts, matchmaking mechanisms are used to automaticly make teams for competitive video games. Traditionally the objective of such systems is to create balanced matches, opposing players or teams of a relatively equivalent level (Delalleau et al., 2012; Véron et al., 2014). It would be interesting to integrate such a system with the previously proposed platform. By separating the players on different servers according to their level and making them all play against the same implementation of a bot, it would then be possible to test the bot's adaptability to the different levels of players. The believability being perceived differently depending on the level of the player, it is important that a bot can adapt to his opponent to be believable.

6.3 Publications

The research conducted during this thesis has been published and presented at international conferences as follows:

1. **Even, 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*, pp. 657–668

2. https://github.com/Blizzard/s2client-api

2. **Even, C.** (2018). « Bot Believability Assessment : a Novel Protocol & Analysis of Judge Expertise ». In: *17th International Conference on Cyberworlds (CW)*, pp. 96–101
3. Buche, C. (2018). « Autonomous virtual player in a video game imitating human players: the ORION framework ». In: *17th International Conference on Cyberworlds (CW)*, pp. 108–113

APPENDIX A

A. Unreal Tournament 2004 Tutorial (in French)

Règles et commandes du jeu :

Le jeu :

Unreal Tournament 2004 (UT2004) est un jeu de tir à la première personne (FPS) développé par Epic Games et Digital Extremes. Il est composé de 10 modes de combat. Le mode utilisé pour cette expérience est le "DeathMatch" (= Match à mort). L'objectif du Deathmatch est d'obtenir un maximum de "frags". Un joueur peut gagner un frag en tuant un joueur ennemi. Si un joueur se tue, il perd un frag. Les joueurs peuvent se tuer avec leurs propres armes ou en tombant dans de la lave, des pièges ou des puits sans fond. Le premier joueur à atteindre le nombre de frags limite gagne la partie.

Les commandes :

Clavier :

Z	Avancer
S	Reculer
Q	Gauche
D	Droite
Espace	Sauter
C	S'accroupir

Souris :

①	Tir régulier
②	Tir alternatif
③	Changer d'arme

Lorsque le match commence, tirez pour entrer dans l'arène de combat.

Pour vous déplacer, utiliser les flèches ou les touches Z Q S D.
Pour esquiver ("dodge"), tapez deux fois sur l'une de ces touches pour effectuer un saut rapide dans la direction choisie.

La touche espace permet de sauter. Vous pouvez réaliser un double saut lorsque vous êtes au sommet de votre premier saut, appuyez à nouveau sur espace pour aller plus haut. Cela est utile pour rester en l'air dans le combat ou pour atteindre des bonus hors de portée.

Les armes :

Chaque arme dans Unreal Tournament comporte deux modes de tir appelés régulier et alternatif.
Vous commencez le match avec ces deux armes :

Assault Riffle

Le tir régulier permet de tirer des balles à vitesse moyenne et le tir alternatif permet de lancer une grenade. Maintenez le bouton enfoncé pour choisir la distance de tir et relâchez pour lancer la grenade.

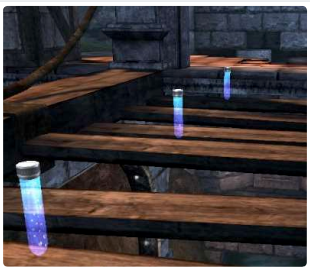
Shield Gun

Le tir régulier est efficace pour les attaques au corps-à-corps. Pour utiliser cette attaque, maintenez le bouton enfoncé pour charger puis approchez-vous de votre ennemi; l'arme lancera l'attaque automatiquement et provoquera d'importants dégâts. Le tir alternatif est un dispositif de protection qui protège partiellement des balles et explosions.

Vous aurez rapidement besoin d'explorer la carte pour trouver d'autres armes plus puissantes. Libre à vous de les chercher et tester leurs deux modes de tir. Gardez en tête que vous devrez aussi trouver des munitions pour ces armes. Elles se trouvent généralement à proximité de l'emplacement des armes.

Les bonus :

L'utilisation de ces objets n'est pas considérée comme de la triche.



Health Vial

Chaque flacon donne 5 points de santé, jusqu'à un maximum de 199.

Health Pack

Réapprovisionne 25 points de santé, jusqu'à un maximum de 100.

Keg O'Health

Rempli le niveau de santé jusqu'à 199 !

Shield Pack

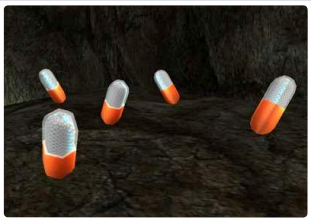
Donne 50 points d'armure. L'armure permet de réduire l'impact des dégâts sur le niveau de santé. Une aura orange apparaît lorsque des dégâts sont subis.

Super Shield Pack

Donne 100 points d'armure, jusqu'à un maximum de 150. Une aura orange apparaît lorsque des dégâts sont subis.

Double Damage

Double la puissance de vos armes pour une durée de 30 secondes. Cet amplificateur de dommages est souvent placé dans un endroit caché ou difficil d'accès. Une aura violette apparaît autour de l'arme.



Adrenaline

Ce bonus sera désactivé pour l'expérience.

Continuer

APPENDIX B

B. Questionnaire for the Experiment No. 1 (in French)

Merci de répondre aux questions suivantes :

Problème technique : ☐

Dans cette partie, la musique était :

Stressante — Relaxante

Par rapport à la partie précédente, le rythme de la musique était :

○ Plus lent ○ Pareil ○ Plus rapide

Pas du tout sûr — Tout à fait sûr

Avez-vous eu le sentiment d'être motivé par la musique ?

Pas du tout — Complètement

En comparant avec la partie précédente, le niveau de votre adversaire était :

○ Nettement moins bon ○ Du même niveau ○ Nettement meilleur

Pas du tout sûr — Tout à fait sûr

D'après vous, votre adversaire était contrôlé par :

Un programme informatique Un humain

Pas du tout sûr — Tout à fait sûr

La durée du match précédent était :

Plus courte De la même durée Plus longue

Pas du tout sûr — Tout à fait sûr

Pensez-vous avoir exploré la totalité de la carte ?

Oui Non

Pas du tout sûr — Tout à fait sûr

Par rapport à la partie précédente, la carte était :

Plus difficile à parcourir Aussi diffcile à parcourir Plus facile à parcourir

Pas du tout sûr — Tout à fait sûr

La position des armes et bonus (Powerups) était :

Aléatoire Prédéfinie

Pas du tout sûr — Tout à fait sûr

Dans cette partie, vous avez trouvez que la carte était :

Trop petite — Trop grande

Continuer

APPENDIX C

C. Final Questionnaire for the Experiment No. 1 (in French)

D'après vous, quel était l'objectif de l'expérience ?

A partir de quel moment (environ) avez-vous compris l'objectif ? (cochez la case)
Après la :
□ 1ère partie □ 2ème □ 3ème □ 4ème □5ème □6ème □7ème □8ème

Avez-vous modifié votre façon de jouer par rapport à cet objectif ? (cochez la case)
□ Oui □ Non

Avez-vous des remarques :

APPENDIX D

D. Questionnaire to Evaluate the Level of Expertise in the Experiment No. 2 (in French)

Merci de répondre aux questions suivantes :

A quelle fréquence jouez vous aux jeux vidéo ?

- ○ Tous les jours
- ○ Plusieurs fois par semaine
- ○ Seulement le weekend
- ○ Quelques fois par mois
- ○ Uniquement pendant les vacances
- ○ Jamais

Quels sont les supports que vous utilisez pour jouer aux jeux vidéo?

- ☐ Ordinateur (jeux sur CD ou sur internet)
- ☐ Console de salon (Xbox, Playstation ou wii par exemple)
- ☐ Console portable (Game Boy, PSP par exemple)
- ☐ Borne d'arcarde (Salle de jeux par exemple)
- ☐ Autre support (Téléphone portable ou baladeur MP3 par exemple)

A quels types de jeux jouez-vous?

Vos choix :

Votre classement :

- First-Person Shooter (expl : counter-Strike)
- Jeux de stratégie (expl : Age of empire)
- Jeux de plateforme (expl : Rayman)
- Jeux d'aventure, d'action (expl : Assassin's Creed)
- Jeux de Rôle ou RPG: Role Playing Game (expl : Final Fantasy)
- Jeux ludo-éducatifs (expl : Adibou)
- Jeux de gestion (expl : Zoo Tycoon)
- Jeux de simulation (expl : Sims)
- Jeux de sports (expl : Fifa, PES)
- Jeux de courses (expl : Grand Turismo, Mario Kart)
- MMORPG = Jeux de rôle massivement multi-joueurs (expl : World of Warcraft)
- Jeux d'activité physique ou sportive (expl : Wii, Kinect, Playstation Move)

Vous jouez :

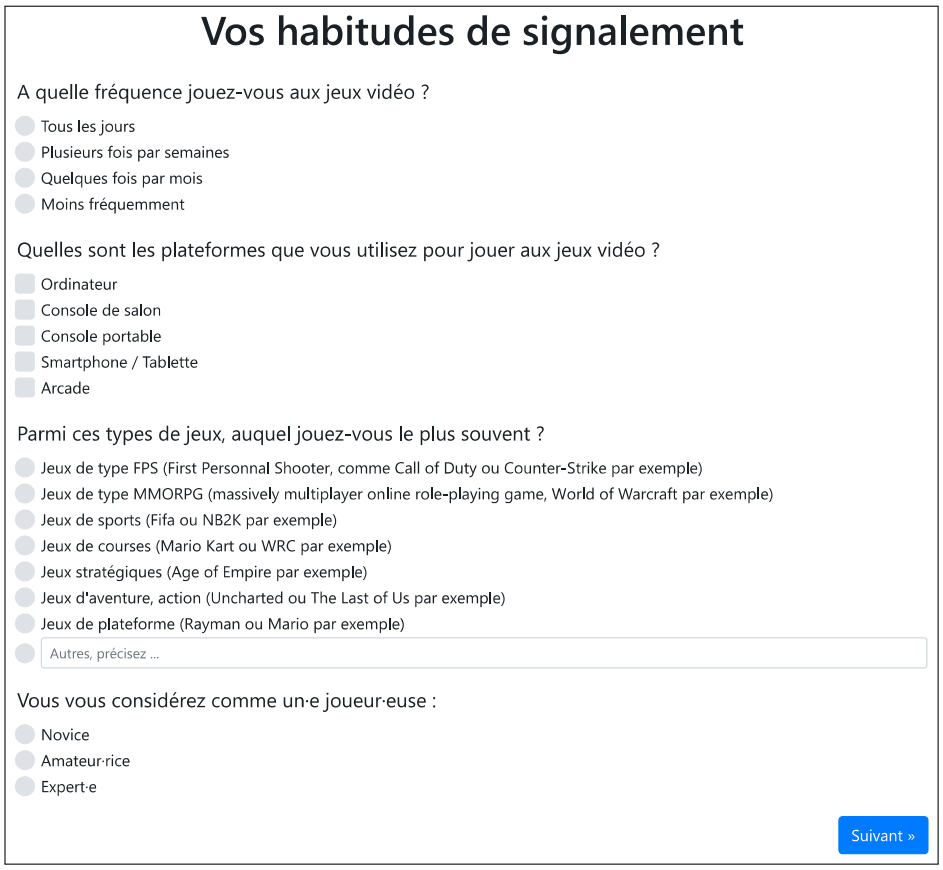
- [ ] Seul(e)
- [ ] Avec des joueurs virtuels
- [ ] En ligne avec des inconnus
- [ ] En ligne avec des amis/famille
- [ ] Avec des joueurs physiquement présents

Continuer

APPENDIX E

E. Material for the Experiment No. 3 (in French)

E.1 Pre-Experiment Questionnaire



Vos habitudes de signalement

A quelle fréquence jouez-vous aux jeux vidéo ?

- Tous les jours
- Plusieurs fois par semaines
- Quelques fois par mois
- Moins fréquemment

Quelles sont les plateformes que vous utilisez pour jouer aux jeux vidéo ?

- Ordinateur
- Console de salon
- Console portable
- Smartphone / Tablette
- Arcade

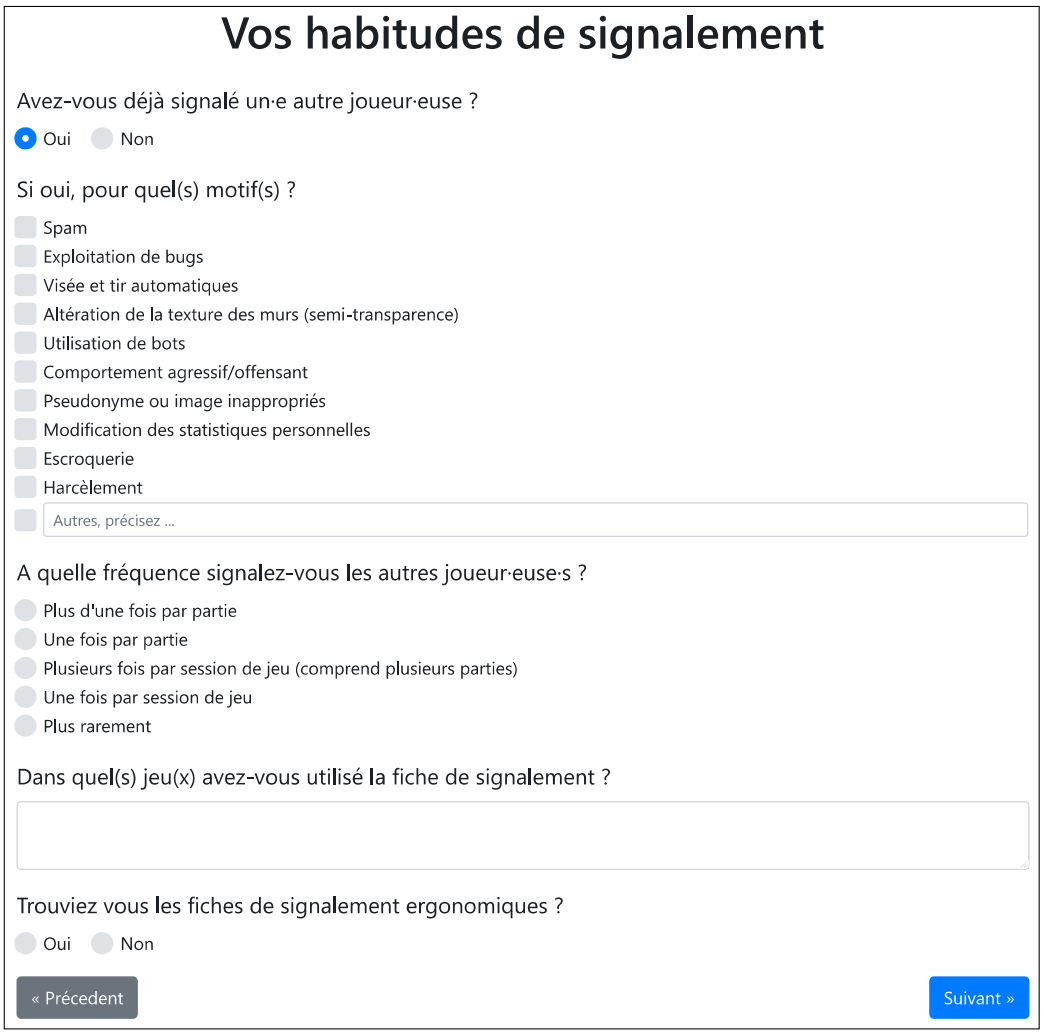
Parmi ces types de jeux, auquel jouez-vous le plus souvent ?

- Jeux de type FPS (First Personnal Shooter, comme Call of Duty ou Counter-Strike par exemple)
- Jeux de type MMORPG (massively multiplayer online role-playing game, World of Warcraft par exemple)
- Jeux de sports (Fifa ou NB2K par exemple)
- Jeux de courses (Mario Kart ou WRC par exemple)
- Jeux stratégiques (Age of Empire par exemple)
- Jeux d'aventure, action (Uncharted ou The Last of Us par exemple)
- Jeux de plateforme (Rayman ou Mario par exemple)
- Autres, précisez ...

Vous vous considérez comme un·e joueur·euse :

- Novice
- Amateur·rice
- Expert·e

Suivant »



Vos habitudes de signalement

Avez-vous déjà signalé un·e autre joueur·euse ?

- (•) Oui
- ( ) Non

Si oui, pour quel(s) motif(s) ?

- [ ] Spam
- [ ] Exploitation de bugs
- [ ] Visée et tir automatiques
- [ ] Altération de la texture des murs (semi-transparence)
- [ ] Utilisation de bots
- [ ] Comportement agressif/offensant
- [ ] Pseudonyme ou image inappropriés
- [ ] Modification des statistiques personnelles
- [ ] Escroquerie
- [ ] Harcèlement
- [ ] Autres, précisez ...

A quelle fréquence signalez-vous les autres joueur·euse·s ?

- ( ) Plus d'une fois par partie
- ( ) Une fois par partie
- ( ) Plusieurs fois par session de jeu (comprend plusieurs parties)
- ( ) Une fois par session de jeu
- ( ) Plus rarement

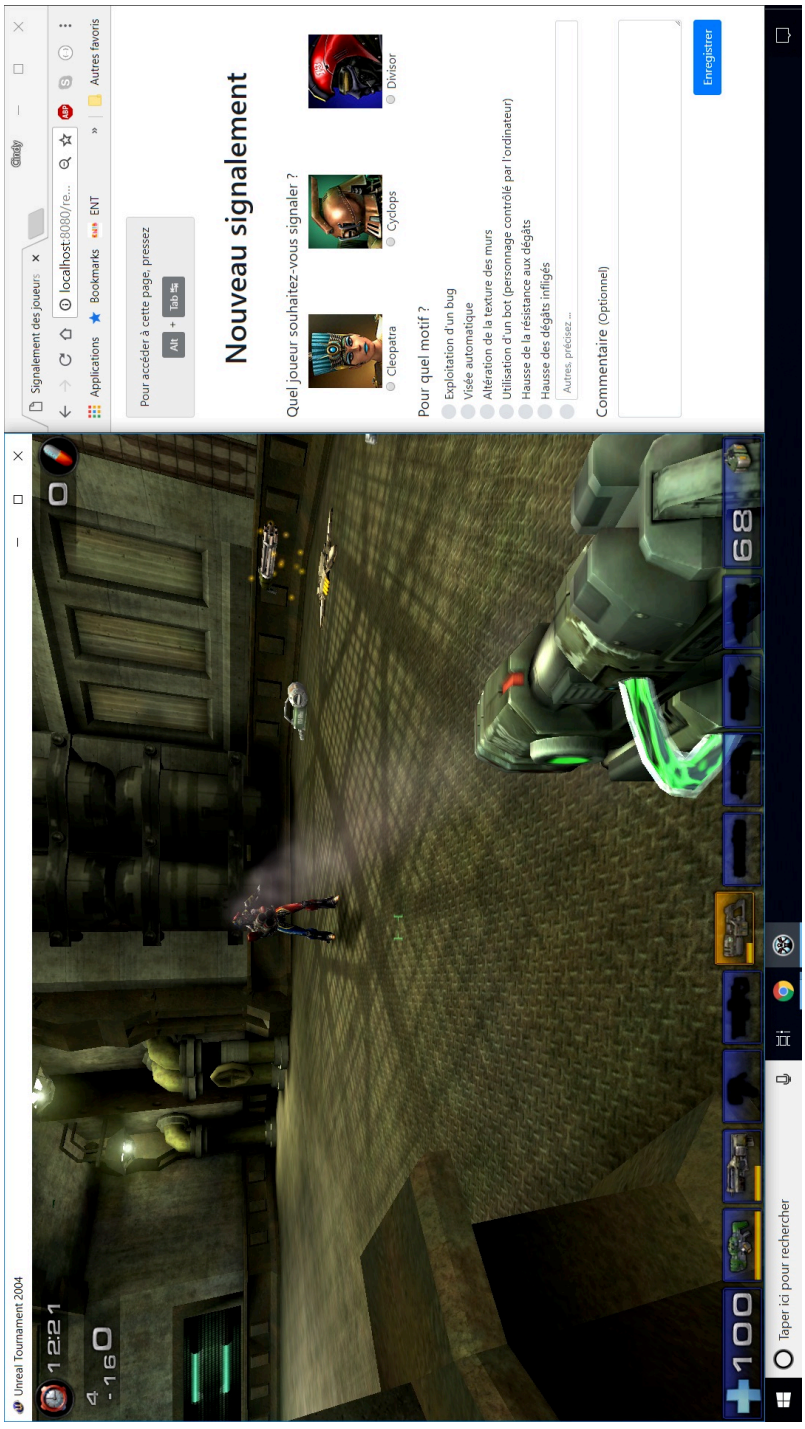
Dans quel(s) jeu(x) avez-vous utilisé la fiche de signalement ?

Trouviez vous les fiches de signalement ergonomiques ?

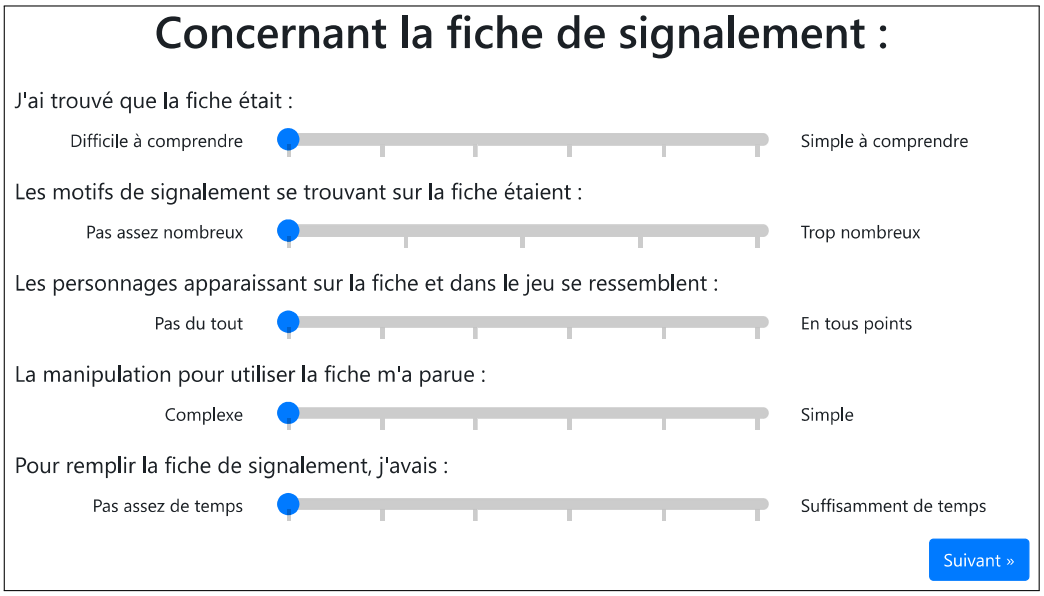
- ( ) Oui
- ( ) Non

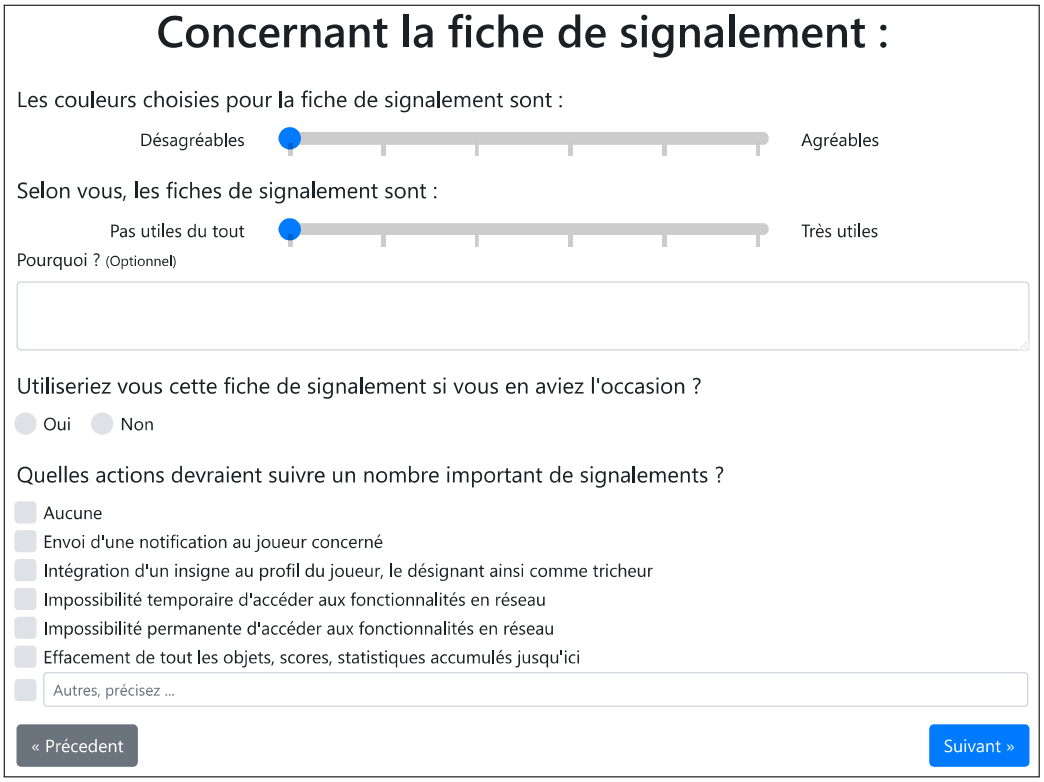
« Précedent Suivant »

E.2 Screenshot of the Experiment in Process



E.3 Post-Experiment Questionnaire





Question sur les adversaires :

Selon vous, combien de fois avez-vous utilisé la fiche pour signaler :

	Jamais	Entre 1 et 3 fois	Entre 4 et 6 fois	Plus de 6 fois
L'exploitation d'un bug				
Une automatisation de la visée et du tir				
Une altération de la texture des murs (se traduisant par une anticipation de votre position)				
L'utilisation d'un bot programmé				
Une hausse de la résistance de l'adversaire				
Une hausse des dégâts produit par une arme de l'adversaire				

De manière générale, les joueurs en face avaient un comportement :

Nuisant à l'expérience de jeu — Améliorant l'expérience de jeu

« Précedent Suivant »

Question sur les adversaires :

Quel joueur avez vous le plus signalé ?

 Tutoriel
 Tutoriel
 Tutoriel

Pourquoi ?

Selon vous, au total :

Combien de joueurs humains avez-vous affronté ?

0 1 2 3 — Pas du tout sûr — Tout à fait sûr

Combien de bots avez-vous affronté ?

0 1 2 3 — Pas du tout sûr — Tout à fait sûr

Les joueurs humains :

Trichaient moins que les bots — Trichaient plus que les bots

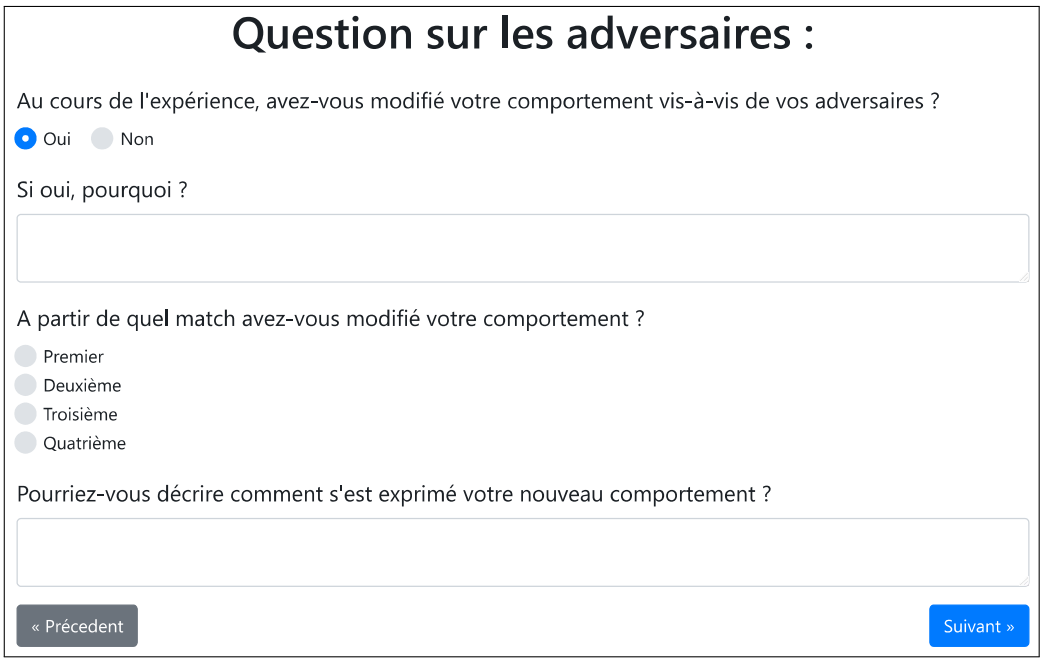
Selon vous, si les joueurs savaient comment tricher sans sanctions, ils :

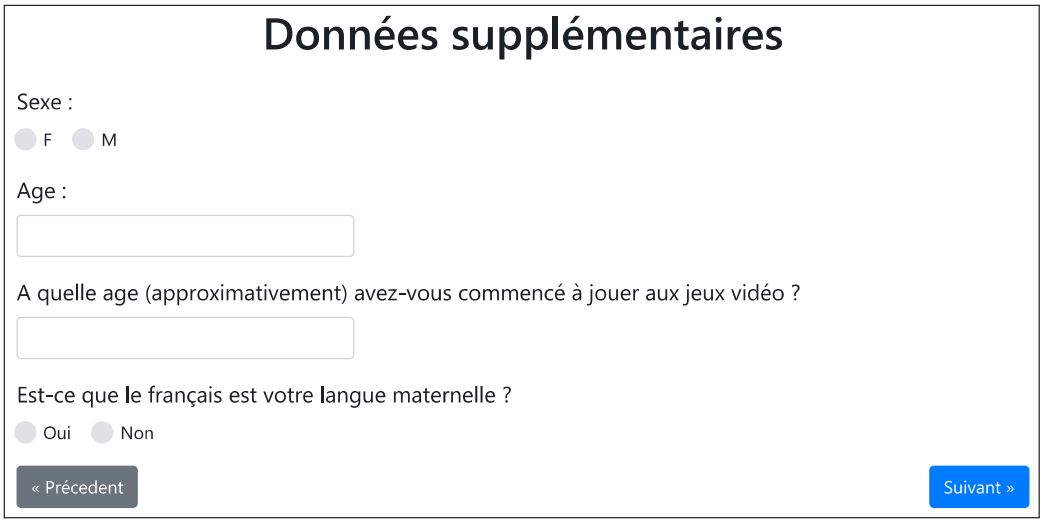
Tricheraient — Ne tricheraient pas

Les bots que vous affrontiez vous paraissaient :

Pas du tout crédible — Très crédible

« Précedent Suivant »


Question sur les adversaires :
Au cours de l'expérience, avez-vous modifié votre comportement vis-à-vis de vos adversaires ?
Oui
Non
Si oui, pourquoi ?
A partir de quel match avez-vous modifié votre comportement ?
Premier
Deuxième
Troisième
Quatrième
Pourriez-vous décrire comment s'est exprimé votre nouveau comportement ?
« Précedent
Suivant »


Données supplémentaires
Sexe :
F
M
Age :
A quelle age (approximativement) avez-vous commencé à jouer aux jeux vidéo ?
Est-ce que le français est votre langue maternelle ?
Oui
Non
« Précedent
Suivant »

BIBLIOGRAPHY

Acampora, G., Loia, V., Vitiello, A., (2012). « Improving game bot behaviours through timed emotional intelligence ». In: *Knowledge-Based Systems* 34, pp. 97–113. ISSN: 09507051 (cit. on pp. 17, 22, 27).

Alayed, H., Frangoudes, F., 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, pp. 1–8 (cit. on pp. 4, 75).

Arrabales, R., Ledezma, A., Sanchis, A., (2010). « ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents ». In: *Journal of Consciousness Studies* 17.3-1, pp. 131–164 (cit. on p. 19).

Arrabales, R., Ledezma, A., Sanchis, A., (2012). « ConsScale FPS: Cognitive Integration for Improved Believability in Computer Game Bots ». In: *Believable Bots: Can Computers Play Like People?* Ed. by Philip Hingston. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 193–214. ISBN: 978-3-642-32323-2 (cit. on p. 19).

Bailenson, J. N., Aharoni, E., Beall, A. C., Guadagno, R. E., Dimov, A., 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*, pp. 216–223 (cit. on p. 14).

Bates, J. (1994). « The role of emotion in believable agents ». In: *Communications of the ACM* 37.7, pp. 122–125. ISSN: 00010782 (cit. on p. 14).

Bernacchia, M., Hoshino, J., (2014). « AI platform for supporting believable combat in role-playing games ». In: *Proceedings of the 19th Game Programming Workshop in Japan*, pp. 139–144 (cit. on p. 21).

Bevacqua, E., Stanković, I., Maatallaoui, A., Nédélec, A., De Loor, P., (2014). « Effects of coupling in human-virtual agent body interaction ». In: *Intelligent Virtual Agents*. Springer, pp. 54–63 (cit. on p. 14).

Bída, M., Černỳ, M., Gemrot, J., Brom, C., (2012). « Evolution of GameBots project ». In: *International Conference on Entertainment Computing*. Springer, pp. 397–400 (cit. on pp. 38, 86, 88).

Bogdanovych, A., Trescak, T., Simoff, S., (2016). « What makes virtual agents believable? » In: *Connection Science*. ISSN: 0954-0091 (cit. on pp. 15, 22).

Bossard, C., Benard, R., De Loor, P., Kermarrec, G., Tisseau, J., (2009). « An exploratory evaluation of virtual football player's believability ». In: *Proceedings of 11th Virtual Reality International Conference (VRIC'09)*, pp. 171–172 (cit. on pp. 22, 27).

Bosse, T., 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, pp. 1–8 (cit. on p. 14).

Brace, I. (2008). *Questionnaire design: How to plan, structure and write survey material for effective market research*. Kogan Page Publishers (cit. on p. 26).

Brand, S. (1999). « Escaping the Digital Dark Age. » In: *Library Journal* 124.2, pp. 46–48 (cit. on p. 7).

Buche, C., **Even, C.**, Soler, J., (2018). « Autonomous virtual player in a video game imitating human players: the ORION framework ». In: *17th International Conference on Cyberworlds (CW)*, pp. 108–113 (cit. on p. 89).

Campbell, D. T., Fiske, D. W., (1959). « Convergent and discriminant validation by the multitrait-multimethod matrix. » In: *Psychological bulletin* 56.2, p. 81 (cit. on p. 71).

Campbell, M., Hoane, A. J., Hsu, F.-h., (2002). « Deep blue ». In: *Artificial intelligence* 134.1, pp. 57–83 (cit. on p. 15).

Chakraborty, J., Norcio, A. F., (2009). « Cross cultural computer gaming ». In: *International Conference on Internationalization, Design and Global Development*. Springer, pp. 13–18 (cit. on p. 21).

Chen, K.-T., Hong, L.-W., (2007). « User identification based on gameplay activity patterns ». In: *Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games*. ACM, pp. 7–12 (cit. on p. 4).

Chen, K.-T., Jiang, J.-W., Huang, P., Chu, H.-H., Lei, C.-L., Chen, W.-C., (2009). « Identifying MMORPG bots: A traffic analysis approach ». In: *EURASIP Journal on Advances in Signal Processing* 2009, p. 3 (cit. on pp. 4, 6).

Cheung, G., Huang, J., (2011). « Starcraft from the stands: understanding the game spectator ». In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 763–772 (cit. on p. 24).

Clark, P., Etzioni, O., (2016). « My Computer is an Honor Student but how Intelligent is it? Standardized Tests as a Measure of AI ». In: *AI Magazine* 37.1, pp. 5–12 (cit. on pp. 9, 31).

Coleridge, S. T. (1817). *Biographiae litteraria of biographical sketches of my literary life and opinions*. Rest Fenner, 23 Patermaster Row (cit. on p. 14).

Collins, J., Hisrt, W., Tang, W., Luu, C., Smith, P., Watson, A., Sahandi, R., (2016). « EDTree: Emotional Dialogue Trees for Game Based Training ». In: *International Conference on Technologies for E-Learning and Digital Entertainment*. Springer, pp. 77–84 (cit. on p. 5).

Consalvo, M. (2009). *Cheating: Gaining advantage in videogames*. Mit Press (cit. on p. 2).

Cronbach, L. J., Meehl, P. E., (1955). « Construct validity in psychological tests. » In: *Psychological bulletin* 52.4, p. 281 (cit. on p. 71).

Cronk, B. C. (2017). *How to use SPSS®: A step-by-step guide to analysis and interpretation*. Routledge (cit. on p. 81).

De Rosis, F., Pelachaud, C., Poggi, I., (2004). « Transcultural believability in embodied agents: a matter of consistent adaptation ». In: *Agent Culture: Human-Agent Interaction in a Multicultural World*, pp. 75–106 (cit. on p. 21).

Delalleau, O., Contal, E., Thibodeau-Laufer, E., Ferrari, R. C., Bengio, Y., Zhang, F., (2012). « Beyond skill rating: Advanced matchmaking in ghost recon online ». In: *IEEE Transactions on Computational Intelligence and AI in Games* 4.3, pp. 167–177 (cit. on p. 88).

Dolnicar, S., Grün, B., Leisch, F., (2011). « Quick, simple and reliable: Forced binary survey questions ». In: *International Journal of Market Research* 53.2, p. 231 (cit. on p. 35).

Elliott, A. C., Woodward, W. A., (2007). *Statistical analysis quick reference guidebook: With SPSS examples*. Sage (cit. on p. 81).

Even, C., Bosser, A.-G., 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*, pp. 657–668 (cit. on pp. 13, 88).

— (2018). « Bot Believability Assessment : a Novel Protocol & Analysis of Judge Expertise ». In: *17th International Conference on Cyberworlds (CW)*, pp. 96–101 (cit. on pp. 51, 89).

Ferranti, (May 5, 1951). *Faster than Thought: The Ferranti Nimrod Digital Computer*. URL: http://www.goodeveca.net/nimrod/NIMROD_Guide.html (cit. on p. 2).

Friedman, H. H., Amoo, T., (1999). « Rating The Rating Scales ». In: *The Journal of Marketing Management* 9.3, pp. 114–123. ISSN: 10711988 (cit. on p. 29).

Gemrot, J., Kadlec, R., Bída, M., Burkert, O., Píbil, R., Havlíček, J., Zemčák, L., Šimlovič, J., Vansa, R., Štolba, M., (2009). « Pogamut 3 can assist

developers in building AI (not only) for their videogame agents ». In: *Agents for games and simulations*. Springer, pp. 1–15 (cit. on p. 38).

Ghasemi, A., Zahediasl, S., (2012). « Normality tests for statistical analysis: a guide for non-statisticians ». In: *International journal of endocrinology and metabolism* 10.2, p. 486 (cit. on p. 81).

Gilovich, T., Griffin, D., Kahneman, D., (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press (cit. on p. 33).

Goffman, E. (1963). *Behavior in public places: Notes on the social organization of gatherings*. Free Press New York (cit. on p. 14).

Goldsmith Jr. T. T., Mann, E. R., (Dec. 17, 1948). « Cathode-ray tube amusement device ». 2455992 (US) (cit. on p. 1).

Gorman, B., Thurau, C., Bauckhage, C., Humphrys, M., (2006). « Believability Testing and Bayesian Imitation in Interactive Computer Games ». In: *From Animals to Animats 9* 1, pp. 655–666. ISSN: 10514651 (cit. on pp. 13, 19, 22, 24, 29).

Heeks, R. (2010). « Understanding G̈old Farmingänd Real-Money Trading as the Intersection of Real and Virtual Economies ». In: *Journal For Virtual Worlds Research* 2.4. ISSN: 1941-8477. DOI: 10.4101/jvwr.v2i4.868. URL: https://journals.tdl.org/jvwr/index.php/jvwr/article/view/868 (cit. on p. 4).

Heeter, C. (1992). « Being there: The subjective experience of presence ». In: *Presence: Teleoperators & Virtual Environments* 1.2, pp. 262–271 (cit. on p. 14).

Hilaire, S., Kim, H.-c., Kim, C.-k., (2010). « How to deal with bot scum in MMORPGs? » In: *Communications Quality and Reliability (CQR), 2010 IEEE International Workshop Technical Committee on*. IEEE, pp. 1–6 (cit. on p. 4).

Hingston, P. (Sept. 2009). « A Turing Test for Computer Game Bots ». In: *IEEE Transactions on Computational Intelligence and AI in Games* 1.3, pp. 169–186. ISSN: 1943-068X (cit. on pp. 15, 16, 22, 23, 27–29, 34–36, 69).

— (Aug. 2010). « A new design for a Turing Test for Bots ». In: *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*. IEEE, pp. 345–350. ISBN: 978-1-4244-6295-7 (cit. on pp. 16, 22, 36).

Hinkkanen, T., Kurhila, J., Pasanen, T. A., (2008). « Framework for evaluating believability of non-player characters in games ». In: *AI and Machine Consciousness* (cit. on pp. 19, 20).

Hoorn, N., Togelius, J., Wierstra, D., Schmidhuber, J., (May 2009). « Robust player imitation using multiobjective evolution ». In: *2009 IEEE*

Congress on Evolutionary Computation. IEEE, pp. 652–659. ISBN: 978-1-4244-2958-5 (cit. on p. 13).

Hyman, H. H., Center, N. O. R., (1954). *Interviewing in social research*. A research project of the National Opinion Research Center. University of Chicago Press (cit. on p. 26).

Jackson, L. A., Wang, J.-L., (2013). « Cultural differences in social networking site use: A comparative study of China and the United States ». In: *Computers in Human Behavior* 29.3, pp. 910–921. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2012.11.024 (cit. on p. 23).

Jin, G. (2006). « Chinese Gold Farmers in the Game World ». In: *Consumers, Commodities & Consumption*. Vol. 7. 2. Consumers Studies Research Network. URL: http://csrn.camden.rutgers.edu/newsletters/7-2/jin.htm (cit. on p. 4).

Kaytoue, M., Silva, A., Cerf, L., (2012). « Watch me playing, i am a professional: a first study on video game live streaming ». In: *Proceedings of the 21st international conference companion on World Wide Web* June 2009, pp. 1181–1188 (cit. on p. 24).

Koehler, D. J., Harvey, N. E., (2004). *Blackwell handbook of judgment and decision making.* Blackwell Publishing. ISBN: 978-1-405-10746-4 (cit. on p. 33).

Krosnick, J. A. (2002). « The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be ». In: *Survey nonresponse*, pp. 87–100 (cit. on p. 35).

Kuny, T. (1998). « The digital dark ages? Challenges in the preservation of electronic information ». In: *International preservation news* 17, pp. 8–13 (cit. on p. 7).

Laird, J. E., 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*, pp. 54–58 (cit. on pp. 13, 19, 22, 24, 27).

Le Hy, R., Arrigoni, A., Bessière, P., Lebeltel, O., (June 2004). « Teaching Bayesian behaviours to video game characters ». In: *Robotics and Autonomous Systems* 47.2-3, pp. 177–185. ISSN: 09218890 (cit. on p. 13).

Lee, Y.-H., Wohn, D. Y., (2012). « Are there cultural differences in how we play? Examining cultural effects on playing social network games ». In: *Computers in Human Behavior* 28.4, pp. 1307–1314. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2012.02.014 (cit. on p. 21).

Lim, S., Reeves, B., (2010). « Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player ». In: *International Journal of Human Computer Studies* 68.1-2, pp. 57–68 (cit. on pp. 9, 28).

Livingstone, D. (Jan. 2006). « Turing's test and believable AI in games ». In: *Computers in Entertainment* 4.1, p. 6. ISSN: 15443574 (cit. on pp. 13, 20, 24, 27, 29).

Llargues Asensio, J. M., Peralta, J., Arrabales, R., Bedia, M. G., Cortez, P., Peña, A. L., (2014). « Artificial Intelligence approaches for the generation and assessment of believable human-like behaviour in virtual characters ». In: *Expert Systems with Applications* 41.16, pp. 7281–7290. ISSN: 09574174 (cit. on pp. 13, 22, 24, 26).

Loyall, A. B. (1997). « Believable Agents: Building Interactive Personalities ». PhD thesis. Carnegie Mellon University. ISBN: 0-89791-877-0 (cit. on pp. 8, 15).

Lucas, S. M., Mateas, M., Preuss, M., Spronck, P., Togelius, J., (2012). « Artificial and Computational Intelligence in Games (Dagstuhl Seminar 12191) ». In: *Dagstuhl Reports* 2.5, pp. 43–70. ISSN: 2192-5283 (cit. on p. 13).

Lugrin, B., Frommel, J., André, E., (2017). « Combining a Data-driven and a Theory-based Approach to Generate Culture-dependent Behaviours for Virtual Characters ». In: *Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies*. Ed. by Colette Faucher. Springer (cit. on p. 21).

Mac Namee, B. (Aug. 2004). « Proactive Persistent Agents: Using Situational Intelligence to Create Support Characters in Character-Centric Computer Games ». PhD thesis. University of Dublin, Trinity College (cit. on pp. 13, 20–22, 27, 29).

MacLean, C. L., Dror, I. E., (2016). « A primer on the psychology of cognitive bias ». In: *Blinding as a Solution to Bias*, pp. 13–24 (cit. on p. 33).

Magnenat-Thalmann, N., Kim, H., Egges, A., Garchery, S., (2005). « Believability and Interaction in Virtual Worlds ». In: *Proceedings of the 11th International Multimedia Modelling Conference*. IEEE, pp. 2–9 (cit. on p. 14).

Marcus, G., Rossi, F., Veloso, M., (2016). « Beyond the Turing Test ». In: *AI Magazine* 37.1, pp. 3–4 (cit. on p. 15).

McDonough, J. P., Olendorf, R., Kirschenbaum, M., Kraus, K., Reside, D., Donahue, R., Phelps, A., Egert, C., Lowood, H., Rojo, S., (2010). *Preserving virtual worlds final report*. Tech. rep. (cit. on p. 7).

McGlinchey, S., Livingstone, D., (2004). « What believability testing can tell us ». In: *Proceedings of the International Conference on Computer*

Games: Artificial Intelligence, Design, and Education, pp. 273–277 (cit. on pp. 13, 20, 22).

Mulligan, J., Patrovsky, B., (2003). *Developing online games: An insider's guide*. New Riders (cit. on pp. xi, 7).

Oliveira, M., Henderson, T., (2003). « What online gamers really think of the Internet? » In: *Proceedings of the 2nd workshop on Network and system support for games*. ACM, pp. 185–193 (cit. on p. 6).

Paritosh, P., Marcus, G., (2016). « Toward a Comprehension Challenge, Using Crowdsourcing as a Tool ». In: *AI Magazine* 37.1, pp. 23–30 (cit. on p. 26).

Poggi, I., Pelachaud, C., Rosis, F., Carofiglio, V., De Carolis, B., (2005). « Greta. a believable embodied conversational agent ». In: *Multimodal intelligent information presentation*. Springer, pp. 3–25 (cit. on p. 14).

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, pp. 1–8 (cit. on pp. 13, 17, 27, 33, 38).

Polceanu, M., Mora, A. M., Jimenez, J. L., Buche, C., Fernandez-Leiva, J., (2016). « The Believability Gene in Virtual Bots ». In: *29th International Florida Artificial Intelligence Research Society Conference (FLAIRS'29)*. AAAI Publications. Key Largo, Florida, USA, pp. 346–349 (cit. on p. 23).

Rehm, M., Bee, N., Endrass, B., Wissner, M., André, E., (2007). « Too close for comfort?: adapting to the user's cultural background ». In: *Proceedings of the international workshop on Human-centered multimedia*. ACM, pp. 85–94 (cit. on p. 21).

Rubin, V. L., Camm, S. C., (2013). « Deception in video games: examining varieties of griefing ». In: *Online Information Review* 37.3, pp. 369–387. DOI: 10.1108/OIR-10-2011-0181. eprint: https://doi.org/10.1108/OIR-10-2011-0181. URL: https://doi.org/10.1108/OIR-10-2011-0181 (cit. on p. xi).

Schuemie, M. J., Van Der Straaten, P., Krijn, M., Van Der Mast, C. A. P. G., (2001). « Research on presence in virtual reality: A survey ». In: *CyberPsychology & Behavior* 4.2, pp. 183–201 (cit. on p. 14).

Scott, B. (2002). « AI game programming wisdom ». In: *The Illusion of Intelligence*, pp. 16–20 (cit. on p. 13).

Shaker, N., Togelius, J., Yannakakis, G. N., Poovanna, L., Ethiraj, V. S., Johansson, S. J., Reynolds, R. G., Heether, L. K., Schumann, T., Gallagher, M., (2013). « The turing test track of the 2012 Mario AI Championship: Entries and evaluation ». In: *IEEE Conference on Compu-*

tational Intelligence in Games (CIG'13). IEEE, pp. 1–8. ISBN: 978-1-4673-5311-3 (cit. on pp. 15, 22, 24, 25, 27).

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., (2016). « Mastering the game of Go with deep neural networks and tree search ». In: *Nature* 529.7587, pp. 484–489 (cit. on p. 15).

Sjöblom, M., Hamari, J., (2017). « Why do people watch others play video games? An empirical study on the motivations of Twitch users ». In: *Computers in Human Behavior* 75, pp. 985–996. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2016.10.019 (cit. on p. 24).

Soni, B., Hingston, P., (June 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)*, pp. 363–369. ISBN: 9781424418213 (cit. on pp. 9, 13, 25).

Tencé, F., Buche, C., (2008). « Automatable Evaluation Method Oriented toward Behaviour Believability for Video Games ». In: *International Conference on Intelligent Games and Simulation (GAMEON'08)*, pp. 39–43. arXiv: arXiv:1009.0501v1 (cit. on p. 19).

Tencé, F., Buche, C., De Loor, P., 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)*. Ed. by Wenji Mao and Lode Vermeersch. Eurosis, pp. 38–45. ISBN: 9789077381540. arXiv: arXiv:1009.0451v1 (cit. on pp. 8, 15, 23).

Tencé, F., Gaubert, L., Soler, J., De Loor, P., Buche, C., (2013). « CHAMELEON: Online Learning for Believable Behaviors Based on Humans Imitation in Computer Games ». In: *Computer Animation and Virtual Worlds (CAVW)* 24.5, pp. 477–496. ISSN: 15464261, 1546427X (cit. on p. 13).

Thawonmas, R., Murakami, S., Sato, T., (2011). « Believable judge bot that learns to select tactics and judge opponents ». In: *IEEE Conference on Computational Intelligence and Games (CIG'11)*, pp. 345–349 (cit. on pp. 16, 33).

Thomas, F., Johnston, O., (1981). *Disney animation: The illusion of life*. Vol. 6. Abbeville Press New York (cit. on p. 14).

Tian, H., Brooke, P. J., Bosser, A.-G., (2012). « Behaviour-Based Cheat Detection in Multiplayer Games with Event-B ». In: *Integrated Formal Methods*. Ed. by John Derrick, Stefania Gnesi, Diego Latella, and Helen Treharne. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 206–220. ISBN: 978-3-642-30729-4 (cit. on p. 4).

Togelius, J. (Mar. 2016). « How to Run a Successful Game-Based AI Competition ». In: *IEEE Transactions on Computational Intelligence and AI in Games* 8.1, pp. 95–100. ISSN: 1943-068X (cit. on p. 15).

Togelius, J., Yannakakis, G. N., Karakovskiy, S., Shaker, N., (2012). « Assessing Believability ». In: *Believable Bots: Can Computers Play Like People?* Ed. by Philip Hingston. Springer Berlin Heidelberg. Chap. 9, pp. 215–230. ISBN: 9783642323232 (cit. on pp. 14, 23, 25, 28, 29).

Turing, A. M. (1950). « Computing machinery and intelligence ». In: *Mind* 59.236, pp. 433–460 (cit. on p. 15).

UFC-Que Choisir, (Sept. 26, 2017). *Transition vers le très haut débit. L'inadmissible amplificateur de la fracture numérique !* URL: https://www.quechoisir.org/action-ufc-que-choisir-transition-vers-le-tres-haut-debit-l-inadmissible-amplificateur-de-la-fracture-numerique-n46732/ (visited on 08/17/2018) (cit. on p. 6).

Verhagen, H., Eladhari, M. P., Johansson, M., McCoy, J., (2013). « Social believablility in games ». In: *Advances in Computer Entertainment: 10th International Conference, ACE 2013, Boekelo, The Netherlands, November 12-15, 2013. Proceedings*. Ed. by Dennis Reidsma, Haruhiro Katayose, and Anton Nijholt. Boekelo, The Netherlands: Springer International Publishing, pp. 649–652 (cit. on pp. 14, 15).

Véron, M., Marin, O., Monnet, S., (2014). « Matchmaking in multi-player on-line games: studying user traces to improve the user experience ». In: *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop*. ACM, p. 7 (cit. on p. 88).

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J., (2017). « Starcraft ii: A new challenge for reinforcement learning ». In: *arXiv preprint arXiv:1708.04782* (cit. on p. 88).

Warpefelt, H., Verhagen, H., (2017). « A model of non-player character believability ». In: *Journal of Gaming & Virtual Worlds* 9.1, pp. 39–53 (cit. on p. 4).

Weibel, D., Wissmath, B., Habegger, S., Steiner, Y., Groner, R., (Sept. 2008). « Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment ». In: *Computers in Human Behavior* 24.5, pp. 2274–2291. ISSN: 07475632 (cit. on pp. 9, 28).

Widaman, K. F. (2018). *Objective Measurement of Subjective Phenomena*. Ed. by Office of Behavioral and National Institutes of Health Social Sciences Research. URL: http://www.esourceresearch.org/eSourceBook/ObjectiveMeasurementofSubjectivePhenomena/

1LearningObjectives/tabid/693/Default.aspx (visited on 11/11/2018) (cit. on p. 71).

Yampolskiy, R. V., Govindaraju, V., (2008). « Embedded noninteractive continuous bot detection ». In: *Computers in Entertainment (CIE)* 5.4, p. 7 (cit. on p. 4).

Yan, J., 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*. ACM, pp. 1–9 (cit. on p. 75).

— (2009). « An investigation of cheating in online games ». In: *IEEE Security & Privacy* 7.3 (cit. on p. 3).

Yannakakis, G. N., Hallam, J., (2009). « Real-time game adaptation for optimizing player satisfaction ». In: *IEEE Transactions on Computational Intelligence and AI in Games* 1.2, pp. 121–133 (cit. on p. 29).

Yannakakis, G. N., Martínez, H. P., (2015). « Ratings are Overrated! » In: *Frontiers in ICT* 2.July, p. 5. ISSN: 2297-198X. DOI: 10.3389/fict.2015.00013. URL: http://journal.frontiersin.org/Article/10.3389/fict.2015.00013/abstract (cit. on p. 35).

Yeung, S. F., Lui, J. C. S., (Sept. 2008). « Dynamic Bayesian approach for detecting cheats in multi-player online games ». In: *Multimedia Systems* 14.4, pp. 221–236. ISSN: 1432-1882. DOI: 10.1007/s00530-008-0113-5. URL: https://doi.org/10.1007/s00530-008-0113-5 (cit. on p. 3).

Zander, S., Armitage, G., (2004). « Empirically Measuring the QoS Sensitivity of Interactive Online Game Players ». In: *Australian Telecommunications Networks & Applications Conference 2004 (ATNAC2004)*, pp. 511–518 (cit. on p. 6).

UNIVERSITE BRETAGNE LOIRE MATHSTIC

ÉCOLE NATIONALE D'INGÉNIEURS DE BREST

Titre : Proposition d'un Protocole et d'un Outil Informatique pour l'Évaluation de la Crédibilité des Joueurs Virtuels dans les Jeux Vidéo Multi-joueurs

Mots clés : Joueur virtuel, bot, jeu vidéo, évaluation, intelligence artificielle

Résumé : L'objectif de cette thèse est de fournir une solution permettant d'évaluer la crédibilité des joueurs virtuels dans les jeux vidéo multi-joueurs. L'état de l'art a permis d'analyser les protocoles existants et ainsi d'identifier sept caractéristiques qui varient considérablement d'une évaluation à l'autre. Notre analyse a également mis en évidence que les méthodes d'évaluation modifient fréquemment le gameplay, introduisant un risque important de biais. Il s'agit là d'une grave lacune car les joueurs virtuels sont ainsi évalués dans un contexte spécifique et non dans le contexte du jeu tel qu'il devrait être joué : cela pourrait fausser les résultats de l'évaluation. À la suite de nos observations, nous avons construit pas à pas un nouveau protocole que nous avons fait évoluer par essais-erreurs en s'appuyant sur les succès de son prédécesseur tout en éliminant les échecs. Pour faciliter la mise en oeuvre de ces essais, nous avons développé un système informatique qui automatise partiellement l'exécution du processus d'évaluation. Ce système est flexible et a montré sa généricité en étant utilisé dans plusieurs configurations. Enfin, nous sommes arrivés à une nouvelle proposition qui permet aux joueurs humains d'évaluer indirectement la crédibilité des joueurs virtuels en utilisant les systèmes de signalement traditionnellement utilisés pour signaler les tricheries, les abus et le harcèlement dans les jeux vidéo en ligne. Le but de notre proposition est d'ajouter des options dans les formulaires de signalement pour signaler la présence de bots. Nous avons émis l'hypothèse que plus un bot est signalé, moins il est crédible. Afin de valider notre approche, nous avons mené une expérience qui a donné des résultats très prometteurs.

Title: Proposal of a Protocol and a Computer Tool for Assessing the Believability of Virtual Players in Multiplayer Video Games

Keywords: Virtual player, bot, video game, assessment, artificial intelligence

Abstract: The objective of this thesis is to provide a solution for assessing the believability of virtual players in multiplayer video games. The literature review allowed us to analyse the existing protocols and thus to identify seven characteristics that vary significantly from one assessment to another. Our analysis also highlighted that evaluation methods frequently modify the gameplay, introducing a significant risk of bias. This is a serious shortcoming as virtual players are thus assessed in a specific context and not in the context of the game the way it should be played: this could skew the results of the assessment. We consequently embarked on a system of trial and error, each new protocol drawing on the successes of its predecessor whilst eliminating the failures. To facilitate the implementation of these trials, we have developed a computer system that partially automates the execution of the evaluation process. This system is flexible and has shown its genericity by being used in several configurations. Finally, we arrived at a novel proposal which allows gamers to indirectly asses the believability of virtual players by using the reporting systems traditionally used to report cheating, abuse and harassment in online video games. The goal of our proposal is to add options in reporting forms that would report the presence of bots. We hypothesised that the more often a bot is reported, the less believable it is. In order to validate our approach, we conducted an experiment which gave very promising results.