Chapter 1
Assessing Believability

Julian Togelius[1], Georgios N. Yannakakis[1],
Sergey Karakovskiy[2], Noor Shaker[1]

Abstract We discuss what it means for a non-player character (NPC) to be believable or human-like, and how we can accurately assess believability. We argue that participatory observation, where the human assessing believability takes part in the game, is prone to distortion effects. For many games, a fairer (or at least complementary) assessment might be made by an external observer that does not participate in the game, through comparing and ranking the performance of human and non-human agents playing a game. This assessment philosophy was embodied in the Turing Test Track of the recent Mario AI Championship, where non-expert bystanders evaluated the human-likeness of several agents and humans playing a version of *Super Mario Bros*. We analyze the results of this competition. Finally, we discuss the possibilities for forming models of believability and of maximizing believability through adjusting game content rather than NPC control logic.

1.1 Introduction

What exactly is believability in a game, what is it good for and how can it be achieved? These are complex but important questions that we will not claim to be able to fully answer. However, in this chapter we will address all these questions from the perspective of *believability assessment*, i.e. how we can accurately judge the believability of a game character. As believability is a fundamentally phenomenological construct, we believe an assessment-based perspective to be appropriate for shedding light on the nature of believability.

[1]IT University of Copenhagen, Rued Langaards Vej 7, 2300 Copenhagen, Denmark
· [2]St. Petersburg State University, Universitetskii prospekt 35, Petergof, Saint-Petersburg 198504, Russia
{juto, yannakakis, nosh}@itu.dk, sergey@marioai.com

In the following, we analyze believability in the context of games, and reason about the complexity of assessing believability and the concerns behind different approaches. We argue that in many cases, believability is better assessed from a third-person perspective rather than a first-person perspective, i.e. where the assessor is not a participant in the game. As an example of third-person believability assessment, we report the results of the Mario AI Turing Test, which was carried out during the Asia Games Show (in collaboration with the IEEE Games Innovation Conference) using uninformed observers as assessors. The discussion about assessing believability is then used to inform a further discussion about how to achieve believability; we propose a solution based on modelling and optimization, analogous to an approach that has been used successfully to maximize player experience. Throughout the chapter we will illustrate the concepts we discuss through how they apply to *Super Mario Bros*, Nintendo's seminal platform game from 1985.

1.2 What is believability?

As far as we can tell, there is no generally agreed or precise definition of believability. Instead, we have a family of related meanings denoted by the same word, somewhat similar to the situation for the word "intelligence". In trying to identify these meanings, we can start with the obvious linguistic fact that believability means that something can be believed by someone. As we are talking about believability in the rather restricted domain of computer game bots or characters, we can add that something about the character can be believed by someone. We might constrain ourselves further by adding the "is real" or "is plausible" to the equation, so that we get "someone believes that some character or bot is real". This leaves us with two broad classes of examples:

- *Character believability*: Someone believes that the character/bot *itself* is real, i.e. an actual living being (or actual autonomous robot etc.)
- *Player believability*: Someone believes that the *player* controlling the character/bot is real, i.e. that a human is playing as that character instead of the character being computer-controlled.

Character believability implies a very high degree of realism; characters can be "realistic" in certain respects (such as textures, movement patterns and dialogue) without having any notable character believability. Viewed this way, character believability currently seems to be reserved for big-budget non-interactive movies (e.g. Hollywood productions such as *Lord of the Rings* where certain entirely computer-generated characters look as real as if they had been human actors in a costume. Hardly any games, or interactive media of any sort, can reasonably aspire to character believability within this technical generation. On the way towards true character believability lurks

the well-known problem of the *uncanny valley*, much discussed in humanoid robotics research: almost, but not completely, believable characters tend to be "creepy" [1] and elicit negative emotions in humans. A number of studies has investigated and, in part, confirmed the uncanny valley theory in virtual characters within games [2, 3] while a large volume of work has focused on the impact of virtual character detail on players' self-reported presence [4] or observers' impressions of character facial images [5]. We do not know whether the uncanny valley phenomenon exists for player believability as well, and there have to our knowledge not been any thorough investigations of this, but some arguments have been made that the phenomenon does indeed exist [6]. The discussion in the rest of the paper will chiefly apply to player believability rather than character believability.

Player believability presupposes that the observer knows that the character on the screen is not real — that it is just a graphical representation of digital processes inside a computer. (This is certainly the case in Super Mario Bros.) Importantly, this means that most aspects of animation and graphics rendering are unimportant for player believability. However, the observer believes that a human has an ongoing input to and (at least partial) control over these processes, and that the human's control is interactive in the sense that the human is aware of what the character is doing in the game.

There are many games and bots that are not designed to be believable, neither in the sense of character believability nor player believability. To begin with, there are game genres completely devoid of NPCs (e.g. puzzle games). There are also games where characters are intended to be predictable and "robotic" in their actions, for example many shooting gallery games and platform games. For example, in Super Mario Bros the most common NPC creatures move with constant speed along a platform and reverse direction or simply fall down when reaching the end of the platform. In other games, there is little chance for the player to assess believability because each NPC is typically only encountered for a few seconds, and then in the middle of a rather chaotic situation; this is the case for the single-player campaign of many FPS games. In general, the game design has considerable influence on the believability of NPCs: either the design showcases the bot and its AI or the game is designed so that the bot's stupidity (or non human-likeness) is absorbed.

Even when the player has a fair chance to observe an NPC for some time and assess its believability, it is clear that player believability is harder to achieve in some games than others. One important factor for the difficulty of achieving player believability is the amount and type of information required from the player (human or algorithm) to effectively play the game. As an example of the importance of the amount of information, chess only requires a few bits of information transferred from the player every turn to select which piece to move to which position (a low-bandwidth channel in the vocabulary of information theory) whereas a first-person shooter such as *Unreal Tournament* requires the player to continuously move the mouse with one hand and

tap keys on the keyboard with the other, often performing several actions per second. The much higher communication bandwidth between player and game goes some way to explain the relative difficulty of creating believable bots for FPS games. In terms of control bandwidth, Super Mario Bros has more in common with the typical FPS than with a board game; although relatively few actions are available to the player at each moment, a player during normal gameplay presses more than one key a second.

As an example of the influence of the type of information required, creating a believable bot for a complex strategy game such as *Civilization* is hard, although there are moderately successful attempts; however, creating a believable bot for a digital version of the strategy game *Diplomacy* is orders of magnitude harder, even though the action space of the latter game is smaller [7]. The key difference here is that text-based (or verbal) communication between players is crucial in Diplomacy, whereas it is often not even possible in Civilization. In terms of the type of information supplied by the player, Super Mario Bros should be relatively easy to achieve character believability in: the player controls the character through a combination of five discrete button presses (left, right, down, A, B). No text input nor continuous input is possible.

Another important question is "believable for whom"? In general, an experienced player will have a much easier time distinguishing between a human-controlled and a computer-controlled character. This is due both to the experienced player having better knowledge about the rules and possible actions in this particular game, or in games in general, and to the experienced player having a larger (though often implicit) knowledge of the patterns of actions exhibited by the artificial intelligence routines found in the game at hand, or in games in general.

It is clear that non-player characters with player believability can bring major advantages for a game. Many games become much more engaging for players who believe that they are playing against fellow human players. Several authors have argued that the appearance of human intelligence and overall human-likeliness that adds value to a computer controlled character and overall, increases the quality of gameplay [8, 9]. There seems to be multiple reasons for why games against humans become more engaging, including the belief that humans are less predictable than computers, that the player through gameplay is capable of inflicting real joy or disappointment in other humans, and the sense of having company in what one does.

Believability — similar to emotions such as frustration and cognitive processes such as attention — is an artificial construct with fuzzy boundaries. These properties make the assessment and computational modelling of believability far from trivial. Before discussing the complex nature of detecting and assessing believability in Section 1.3 we will briefly analyze the relationship between player emotions and believability.

1.2.1 Believability, Player Emotions and Cognition

Game believability is a critical subcomponent of player experience. It can be linked to a stream of player emotions, which may be active simultaneously, usually triggered by events occurring during gameplay but also related to cognitive and behavioural process during gameplay. Games incorporating believable elements, such as bot behaviour, can elicit particular emotional responses to a player which in turn may affect the player's attention level and gaze patterns, reflect on the player's facial expression and even cause bodily alterations (i.e. the player's physiology).

Research in game artificial intelligence is based on several empirical assumptions about believability, human cognition, human-machine interaction, player satisfaction and fun. The primary hypothesis of most studies in the literature is that the generation of believable, human-like opponents [10] leads to increased player enjoyment. While there are indications to support such a hypothesis (e.g. the vast number of multi-player on-line games played) and research endeavors to investigate the relationship between believability of non-player characters (NPCs) and satisfaction of the player [11], there has been no clear evidence that human-like NPC behaviours generate more appealing games.

We need to make clear that we do not necessarily assess player satisfaction or other emotional states by assessing believability; the relationship between the two is complicated, and while it would be interesting and important to clarify it, this is beyond the scope of the current paper.

1.3 Assessing believability

Believability of character behaviour may be viewed as part of player experience. Player experience [12] in general can be measured via reporting (*subjective*); via monitoring a player's physiological responses [13, 14], tracking a player's body, head and facial motion [15] during play (*objective*); or via logging gameplay statistical data that embed behavioural responses of player experience (*gameplay-based*) [16]. On the same basis, one could attempt to adopt any of the three above-mentioned approaches (or even combinations of those) to capture believability within games.

In this paper we focus on the *subjective* approach for assessing believability and we question whether a player can reliably assess believability his/herself and, in part, question existent player testing schemes.

1.3.1 Subjective Assessment

The most direct way to assess believability is to ask the players themselves about their experience when they face or presented with opponents that need to be assessed [17]. Subjective believability assessment can be based on either players' *free response* during play or on *forced* data retrieved through questionnaires.

Naturally, free-response answers may contain richer information about one's believability notion but are hard to analyze appropriately. An experiment designer may decide to annotate the derived text or verbal response into specific critical words or phrases which can then be mapped to believability. However, doing so requires strong assumptions about the validity and the importance of the text/speech clusters identified, and make it hard to automate the assessment. On the other hand, forcing game experiment subjects to report believability through a questionnaire, constraining them into specific questionnaire items, yields data that can be easily used for analysis.

Subjective assessment may yield very accurate models of self-reported believability; however, there are quite a few limitations embedded in this approach. First, there is usually significant experimental noise in the responses of subjects; this may be caused by subject learning effects (the subject might find it easier to spot bots after having seen a few) and self-deception. Second, self-reports can be intrusive if questionnaire items are presented during the gameplay sessions [18, 19]. Third, they are sensitive to subjects' memory limitations if players are asked to express their experience after a lengthy game session (post-experience effect).

Numerous studies have shown that self-reports can guide machine learning algorithms for successfully capturing aspects of player experience in prey/predator [20], physical interactive [21], platform [16, 22] and racing [23] games. We argue that similar approaches can be used for the efficient capture of believability within games.

1.3.2 When to ask?

While efficient methods for minimizing learning effects and self-deception effects have been proposed [24], there is no universally accepted time window within which subjects should be asked to express the level of believability of an NPC. Such a time window should result in a self-reporting process that is both as unobtrusive as possible and suffering from minimal post-experience effects.

Reporting on paper or on a digital questionnaire sheet is the most popular approach to subjective assessment for player experience, either interrupting the subject during gameplay or at the end of a game session. It is straightforward to extend this method to believability assessment. However, a re-

cent attempt on believability assessment, the *2k BotPrize* [25] has focused on making the assessment process part of the game itself. In that study, subjects played a first-person shooter game and were equipped with a special weapon that could be used to distinguish between an AI bot and a human opponent. As a pioneering approach to assessing believability, the BotPrize has received considerable attention, and rightly so. While not directly a subjective approach for believability assessment the innovation of that study is that an in-game element (and not an external questionnaire) defines the platform for the assessment of opponent believability during play. Such an approach initially appears to minimize report intrusiveness. Moreover the ideal interaction time window is set by the player his/herself — the window depends on the interaction time a player spends with particular opponents — bypassing that key protocol design decision.

We argue that in a game setting — such as the FPS scenario of [25] — the believability of opponents cannot be detached from player experience. Thus, assessment during play may turn out to be highly intrusive for both player experience and the assessment of players, while gameplay statistics collected this way may contain experimental artifacts that are difficult to detect and correct for.

Results obtained from the 2010 2k BotPrize during the 2010 IEEE CIG conference corroborate our hypothesis. In particular, it was apparent that some players focused on the task of believability assessment while others focused more on gameplay. Believability assessment done this way introduces a new game mechanic that may appeal to some players. The result was that a human player was characterized as the least human player (even less human than all AI opponents) in the game since he had adopted a strategy seeking to excel to that game mechanic. Some judges raised complaints that the experiment was neither a game (the game was way too intrusive to elicit genuine gameplay experience) nor a proper experimental protocol (since some judges aimed to excel in appearing less believable), but rather a hybrid between the two. Our hypothesis is that during-play believability assessment entails protocol design flaws that do not allow for a reliable evaluation of gameplay believability. Thus, when designing the Turing test track of the Mario AI Championship (our own attempt at believability assessment) we chose to deviate from the approach taken in the BotPrize and use a protocol where judges were asked questions only after a level had been played.

1.3.3 How to ask?

Forced questionnaires could vary from simple tick boxes to multiple choice items. Both the questions and the answers provided could vary from single words to sentences; even though, generally, short and clear question-and-answer items are preferred since lengthy questionnaire items may challenge

short-term memory and cognitive load of the subject. Again taking our cue from methods for assessing player experience, one could identify three types of forced questionnaires for believability assessment:

1. Boolean: subjects have a single boolean answer choice (e.g. *is this believable?*, or, *is this a human playing?*). While this question type is direct and clear, it does not provide with rich information for further analysis.
2. Ranking: subjects are asked to answer questionnaire items given in a ranking/scaling form (e.g. *how believable was that?*). For the use of similar questionnaires in player experience assessment, see [14, 18, 19].
3. Preference: subjects are asked to compare the believability level of two or more sessions of the game (e.g. *which one was more believable?*). For the use of similar questionnaires in player experience assessment, see [26, 24, 23].

There is no single universally accepted approach for questionnaire type even though preliminary results in various user studies suggest that there is an inconsistency between the three questionnaire types when used for assessing player experience. The main disadvantage of ranking questionnaires is that they do not control for the subjective notion of believability across subjects. On the other hand, pairwise preference and boolean questions can minimize subjects' subjective notions of scaling, allow a fair comparison between the answers of different subjects and, thereby, may assist towards a more accurate and subjective capture of believability. For simplicity, we opted for boolean questions in the Mario AI Championship. Being the first time we ran the competition and unfamiliarity with the venue, we decided that boolean questions minimized the risks of technical and/or linguistic glitches.

1.3.4 First person vs. Third Person

Subjective assessment may consider both first person reports (self-reports) but also reports expressed indirectly by experts or external observers. Analogies of this relationship can be found in the comparison between self-expressed experience and annotated experience in affective computing studies [12]. While self-expressed experience comes with several limitations such as self-deception, increased gameplay cognitive and short-term memory load, annotated data (if in large sample sizes) can effectively encapsulate notions of player experience.

In first person assessment, the player is part of gameplay and the same gameplay is the elicitor of his/her player experience. As the player is engaged in playing and forms a vital component of the game-player interaction he/she thereby influences the degree of believability that emerges from that relationship. Thus, believability, gameplay and player experience are interconnected

components of the interactive experience which are hard to separate from each other.

In third person assessment, on the other hand, the player does not play the game his/herself as she is the observer of the playing experience and the gameplay. While one could claim that an observer is not engaged in the *true* experience of gameplay believability (rather in a *quasi*-experience [27]) she is able to concentrate more on the assessment of believability via a higher cognitive focus on the task. Because of the aforementioned limitations of the first person approach our research hypothesis is that third person assessment may lead to more accurate believability assessment. We therefore chose to use third-person assessment in the Mario AI Championship.

It should be added that first-person believability assessment is only possible for games played by at least two players simultaneously. First-person assessment is not possible for a platform game such as Super Mario Bros, at least in the standard version where only a single player character plays at any time, nor for a puzzle game such as *Tetris* or many casual mobile games such as *Angry Birds*. Games for which first-person assessment is possible include FPS games such Unreal Tournament, used in the BotPrize, but also a large variety of strategy games, both real-time (RTS) and turn-based and sports games like racing games, not to mention classic non/digital board games like Chess. Third-person assessment, on the other hand, is possible for all game genres.

As an anecdotal example of the perils of first-person assessment, consider the famous 1997 chess match between reigning world champion Garry Kasparov and the Deep Blue hardware/software. After losing the match, Kasparov complained that the program was playing in a suspiciously human-like manner [28]. It is very likely that the subject could have been affected by one of the many limitations of self-expressed first-person believability assessment, such as self-deception.

1.3.5 Time Required for Assessment

What is the optimal (or minimal) time interval for a user to identify believability? Players in a game have differing perceptual capabilities and cognitive responses. One has to take into account such experimental effects when designing a protocol for assessing believability and to control for them. One way to eliminate the time factor as an artifact from any data collected is to design game sessions that would generate equivalent interaction times with the bot which is under believability assessment. Then players can be asked to express their believability preferences over different game interaction sessions of similar time windows.

The time required to assess believability is also clearly dependent on the game genre. For example, FPS game bots are hard to assess since they often

only appear on screen for a few seconds, and therefore their interaction time with the player is often insufficient [25]. On the other end of the spectrum, opponents in RTS games are observed and interacted with for time windows of several minutes or tens of minutes. Believability is not only dependent on the game genre but also on the surrounding content of the bot; an experiment protocol designer needs to cater to this and control for the game content that is present. That in turn will have an influence in the time required to assess believability appropriately.

In the Mario AI Championship, we chose to let the judges observe the game for a duration of 20-30 seconds, corresponding to the time needed to complete (or fail to complete) a short level, and in the authors' opinion enough to get an idea of the playing style of the player.

1.3.6 The representation of believability

Believability is a conceptual construct in a similar way to any other user, cognitive or affective state with unclear boundaries [29]. Given the fuzzy boundaries of believability and the subjective nature of its notion the representation of it is of key importance: should believability be represented within questionnaires as a state or as a continuous value?

It is even possible that believability is best represented as more than one continuous value. For instance, the emotional dimensions of arousal and valence [30, 31] are used very often to represent emotional states with unclear boundaries in affective computing studies; believability could be represented in a similar fashion.

1.4 The Mario AI Championship: Turing Test Track

The Turing test track of the Mario AI Championship was held during the Asia Games Show 2010[1] in conjunction with the 2010 IEEE Games Innovation Conference[2]. This section presents the competitors, the competition setup and the results obtained.

[1] http://www.asiagameshow.com/

[2] http://ice-gic.ieee-cesoc.org/2010/

1.4.1 The competitors

Five bots and one human player competed in the Turing test track. The five bots were chosen from among the competitors in the Gameplay track of the competition (thus built to play the game as well as possible rather than in a human-like fashion) and the organizers' own experiments. The chosen bots were of varying sophistication, playing strength, and in particular exhibited different playing styles:

- Robin Baumgarten's A* Agent: This agent is based on an A* search algorithm in state space and simulates the future trajectory of both itself and enemy NPCs for each considered actions. This agent runs through the levels, almost continuously jumping and shooting fireballs. Detailed information about this agent can be found in [32].
- Slawomir Bojarski's and Clare Bates Congdon's REALM Agent: A rule-based evolutionary computation agent that evolves rule sets based on abstract vocabulary of conditions and actions with an A* component to determine specific keystrocks for each high level action [33]. This agent exhibits a more human-like behaviour than the other agents by starting to jump before reaching the edge of a gap, attacking and avoiding enemies, grasping power-ups and moving in both directions.
- Forward Agent: A very simple heuristic agent that constantly runs left and jumps when it senses that it is in front of a gap or obstacle.
- Forward Jumping Agent: An even simpler agent, that constantly jumps while running left.
- Erek Speed's Agent: Rule-based controller, evolved with a GA. Maps the whole observation space (22 x 22) onto the action space, resulting in a genome of more than 100 Mb.
- Human player: Along with the five agents, an extra recording from a human player who was not involved in programming any of the bots or the organization of the competition (Nikolay Sohryakov) was used. This human had a skill in the high intermediate range, and a non-exploratory playing style.

1.4.2 Competition organization

Prior to the competition event, videos were recorded of the five AI contestants and the human player playing a short level of the competition version of Super Mario Bros. The videos of gameplay were presented to the audience of the Asia Games Show in a random order and the audience voted on whether the player was a human or an algorithm after each video was completed. Each of the 60 observers was asked to vote whether the Mario they just saw playing was controlled by a computer or a human. The "Not decided" option

Fig. 1.1 Image from the Mario AI: Turing Test Track competition held during the Asia Game Show, 2010.

was also available. Each agent was shown at least twice, and the orderings between the agents were varied so as to prevent order effects. A photograph of the competition presenter and the general competition setup is presented in Figure 1.1.

1.4.3 Competition Results

Table 1.1 presents the final results of the competition. 60 persons attended the competition session and voted. The REALM bot was the winner of the competition since it managed to convince 17 of the observers that it was a human. The REALM bot is the bot that comes closest to the 22 "Human" vote baseline of the human player with only 5 votes away while Erek Speed's bot also did rather well gathering 9 human votes. It is worth noting that the human player got more "computer" votes than the REALM bot (32 and 20, respectively) since the REALM bot left 13 spectators undecided — 7 more than the undecided observers for the human player.

The competition results illustrate the difficulty of assessing believability even in a game such as Super Mario Bros, with low control bandwidth, simple graphics and easy overview of the play area. While the human player got the most of the human votes those were only 22 out of 60. Given this preliminary experimental protocol it appears that the 3rd person assessment approach is appropriate since believability can be successfully assessed; however, results also show the subjective nature of believability and the complexity that arises when one attempts to assess it.

Super Mario Bros Player	Computer	Human	Not decided
Human player	32	22	6
REALM Bot (Evolved ruleset)	30	17	13
Erek Speed (GA)	41	9	10
Robin Baumgarten (A*)	46	6	8
Forward Agent	48	6	6
Forward Jumping Agent	54	0	6

Table 1.1 Turing Test Track competition results

The majority of experiment observers classified the human player as an AI-controlled bot and, a few observers indicated that the A* bot and the forward agent are controlled by humans. While the forward jumping Mario agent does not appear to be believable, the forward moving agent and the A* bot can still mislead a few observers. So, how did this happen? The core mechanics of platform games promote simple forward moving behaviour combined with jumps when necessary; such a playing behaviour is often followed by average players of this game genre. However, a human player with relatively high skills has been used in this experiment, and this might be a possible explanation for misleading it with an AI-controlled bot. On the other hand, the A* bot near-optimal performance resembles the behaviour of very few highly-skilled platform game players. Agents mimicking any of those playing behaviours can apparently mislead a few observers and be assessed as believable. On the contrary, the forward jumping agent does not convince any observer for its human nature since such gameplay is rarely met in humans playing Super Mario Bros.

1.5 From assessing to modelling to optimizing believability

Once we have established a reliable measure of believability, we could use supervised learning techniques to create a model from game configuration to believability. This process would be completely analogous to previous work on computational modelling of player experience [34, 12]: a number of game configurations are presented to a set of users, the users judge their believability, and based on this data set (with believability assessments as target values) a model is inferred that predicts believability based on game configuration. The model could use any of several function representations, for example a multilayer perceptron or a decision tree, and be trained with e.g. neuroevolutionary preference modelling (in case of assessments being expressed as preferences) or more standard supervised learning algorithms (in case of scalar assessments). After this model is obtained, optimization algorithms can be used to tune the game so as to maximize predicted believability. In previous

work, player experience has been optimized for Super Mario Bros through creating a model from in-game player behaviour and level design parameters to predicted player affective states (such as fun and frustration), and new levels thereafter evolved that maximized predicted player experience [16, 35].

A key design question then becomes how to meaningfully parameterize the game configuration, so that the parameters have bearing on believability and create a tractable search space for the optimization algorithm. The obvious candidate would be the control logic for the character that is to be made believable. In Super Mario Bros itself this would be the main character, Mario. A number of good Mario controllers have been developed and submitted the Gameplay track of the Mario AI Championship, including Robin Baumgarten's A*-based controller that won the 2009 edition of the competition, and Slawomir Bojarski's and Clare Bates Congdon's evolutionary rule-based agent that won the 2010 edition, and also won the Turing Test track of the championship as described above. Both of these have several parameters that could conceivably be optimized for believability (and other modifiers could be introduced, such as a probability of "stopping to think" for a while every now and then), given that a good evaluation function was available.

However, it might be worth considering investigating other alternatives as well. In his classic book *The Sciences of the Artificial*, Herbert Simon describes the complex path of an ant walking on the beach, noting that the ant itself to our best evidence has a very simple "control system", before asking whether the apparent complexity of the ant's path is due to the ant itself or the topology and distribution of objects on the beach [36]. Analogously, we may ask to what extent believable behaviour in an algorithm-controlled game agent comes about from the controller and to what extent it is a product of the environment. It seems entirely probable that optimizing the environment for believability, in conjunction with a sufficiently generic NPC controller, could be every bit as effective as optimizing the NPC controller itself. In Super Mario Bros, this could be done through optimizing the design parameters of the levels. It is conceivable that the best effects are reached through combining NPC controller and level design optimization.

1.6 Conclusion

This chapter has discussed a number of aspects of believability from the perspective of believability assessment. We have outlined a number of important choices to consider when assessing believability, and briefly discussed their pros and cons. Throughout the chapter, we have used the platform game Super Mario Bros as a running example, and discussed what it would mean for a Mario player to be believable, and how believability could be achieved and assessed in this context. We reported the design and results of the Turing test track of the 2010 Mario AI Championship, which attempted to mea-

sure believability in this game, taking a number of design choices that differ markedly from the perhaps most well-known attempt at assessing bot believability, the 2k BotPrize. It is clear that there is much research left to do about how believability can be assessed, modelled and optimized, and its relation to other aspects of player experience. We intend to contribute to this discussion through running further iterations of the Mario AI Championship, improving the competition design using lessons we have learned from last year's competition.

Acknowledgements

Thanks to all the participants of the Mario AI Championship: Turing Test Track held in the IEEE GIC conference in Hong Kong, December 2010, sponsored by IDSIA in Lugano. This research was supported in part by the European Union FP7 ICT project *SIREN* (project number 258453) and by the Danish Research Agency project *AGameComIn* (project number 274-09-0083).

References

1. Masahiro, M.: Bukimi no tani (the uncanny valley). Energy (1970)
2. Schneider, E.: Mapping out the uncanny valley: a multidisciplinary approach. In: ACM SIGGRAPH 2008 posters. SIGGRAPH '08, New York, NY, USA, ACM (2008) 33:1–33:1
3. Schneider, E., Wang, Y., Yang, S.: Exploring the Uncanny Valley with Japanese Video Game Characters. In: Proceedings of the DIGRA conference. (2007) 546–549
4. Vinayagamoorthy, V., A, B., Gillies, M., Slater, M., Steed, A.: An Investigation of Presence Response across Variations in Visual Realism. In: Proceedings of the 7th International Conference on Presence. (2004)
5. Seyama, J., Nagayama, R.S.: The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. Presence **16**(4) (2007) 337–351
6. Hayward, D.: Uncanny ai: Artificial intelligence in the uncanny valley. Gamasutra **N/A** (2007)
7. Kemmerling, M., Ackermann, N., Beume, N., Preuss, M., Uellenbeck, S., Walz, W.: Is human-like and well playing contradictory for diplomacy bots? In: Proceedings of the IEEE Symposium on Computational Intelligence and Games. (2009) 209–216
8. Champandard, A.J.: AI Game Development. New Riders Publishing (2004)
9. Bateman, C., Boon, R.: 21st Century Game Design. Charles River Media (2005)
10. Freed, M., Bear, T., Goldman, H., Hyatt, G., Reber, P., Sylvan, A., Tauber, J.: Towards more human-like computer opponents. In: Working Notes of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment. (2000) 22–26
11. Taatgen, N.A., van Oploo, M., Braaksma, J., Niemantsverdriet, J.: How to construct a believable opponent using cognitive modeling in the game of set. In: Proceedings of the fifth international conference on cognitive modeling. (2003) 201–206
12. Yannakakis, G.N., Togelius, J.: Experience-driven Procedural Content Generation. IEEE Transactions on Affective Computing (2011) in print.

13. Yannakakis, G.N., Hallam, J., Lund, H.H.: Entertainment Capture through Heart Rate Activity in Physical Interactive Playgrounds. User Modeling and User-Adapted Interaction, Special Issue: Affective Modeling and Adaptation **18**(1-2) (2008) 207–243
14. Mandryk, R.L., Inkpen, K.M., Calvert, T.W.: Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologies. Behaviour and Information Technology (Special Issue on User Experience) **25**(2) (2006) 141–158
15. Asteriadis, S., Karpouzis, K., Kollias, S.D.: A neuro-fuzzy approach to user attention recognition. In: Proceedings of ICANN, Springer (2008) 927–936
16. Pedersen, C., Togelius, J., Yannakakis, G.N.: Modeling Player Experience for Content Creation. IEEE Transactions on Computational Intelligence and AI in Games **2**(1) (2010) 54–67
17. Hingston, P.: A Turing Test for Computer Game Bots. IEEE Transactions on Computational Intelligence and AI In Games **1**(3) (2009)
18. Drachen, A., Nacke, L., Yannakakis, G.N., Pedersen, A.L.: Correlation between heart rate, electrodermal activity and player experience in First-Person Shooter games. In: In press for SIGGRAPH 2010, ACM-SIGGRAPH Publishers (2010)
19. Pagulayan, R.J., Keeker, K., Wixon, D., Romero, R.L., Fuller, T.: User-centered design in games. The HCI Handbook. Lawrence Erlbaum Associates (2003)
20. Yannakakis, G.N., Hallam, J.: Towards Capturing and Enhancing Entertainment in Computer Games. In: Proceedings of the 4^{th} Hellenic Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence. Volume 3955., Heraklion, Greece, Springer-Verlag (2006) 432–442
21. Yannakakis, G.N., Hallam, J.: Real-time Game Adaptation for Optimizing Player Satisfaction. IEEE Transactions on Computational Intelligence and AI in Games **1**(2) (2009) 121–133
22. Pedersen, C., Togelius, J., Yannakakis, G.N.: Modeling Player Experience in Super Mario Bros. In: Proceedings of the IEEE Symposium on Computational Intelligence and Games, Milan, Italy, IEEE (2009) 132–139
23. Tognetti, S., Garbarino, M., Bonarini, A., Matteucci, M.: Modeling enjoyment preference from physiological responses in a car racing game. In: Proceedings of the IEEE Conference on Computational Intelligence and Games, Copenhagen, Denmark (2010) 321–328
24. Yannakakis, G.N.: Preference Learning for Affective Modeling. In: Proceedings of the Int. Conf. on Affective Computing and Intelligent Interaction, Amsterdam, The Netherlands, IEEE (2009) 126–131
25. Hingston, P.: A New Design for a Turing Test for Bots. In: Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, Copenhagen, Denmark, IEEE (2010) 345–350
26. Yannakakis, G.N., Hallam, J.: Towards Optimizing Entertainment in Computer Games. Applied Artificial Intelligence **21** (2007) 933–971
27. Walton, K.L.: Mimesis as make-believe. Harvard University Press, Cambridge, MA (1990)
28. Kasparov, G.: The chess master and the computer. The New York Review of Books (2010)
29. Calvo, R.A., Mello, S.D.: Affect detection: An interdisciplinary reveiw of models, methods and their applications. IEEE Transactions on Affective Computing **1**(1) (2010) 18–37
30. Feldman, L.: Valence focus and arousal focus: Individual differences in the structure of affective experience. Journal of Personality and Social Psychology **69** (1995) 53–166
31. Russell, J.A.: Core affect and the psychological construction of emotion. Psychological Rev. **110** (2003) 145–172
32. Togelius, J., Karakovskiy, S., Baumgarten, R.: The 2009 Mario AI Competition. In: Evolutionary Computation (CEC), 2010 IEEE Congress on, IEEE (2010) 1–8

33. Bojarski, S., Congdon, C.B.: REALM: A Rule-Based Evolutionary Computation Agent that Learns to Play Mario. In: Proceedings of the IEEE Symposium on Computational Intelligence and Games, Copenhagen, Denmark, IEEE (2010) 83–90
34. Yannakakis, G.N.: How to Model and Augment Player Satisfaction: A Review. In: Proceedings of the 1st Workshop on Child, Computer and Interaction, Chania, Crete, ACM Press (2008)
35. Shaker, N., Yannakakis, G.N., Togelius, J.: Towards Automatic Personalized Content Generation for Platform Games. In: Proceedings of Artificial Intelligence and Interactive Digital Entertainment (AIIDE'10), Palo Alto, CA, AAAI Press (2010) 63–68
36. Simon, H.: The Sciences of the Artificial. MIT Press (1969)