PAGAN for Character Believability Assessment

Cristiana Pacheco
Queen Mary University of London, UK
Email: c.pacheco@qmul.ac.uk
Supervisor: Diego Perez-Liebana, Queen Mary University of London, UK
Host Supervisor: Georgios N. Yannakakis, Institute of Digital Games, Malta
Other collaborators: David Melhart, Antonios Liapis, Institute of Digital Games, Malta

***Abstract*—What makes an agent believable? It depends. The concept of 'believability' is very subjective with researchers developing different definitions every time. As a result, assessing it is not straightforward either. Many have adapted the Turing Test in attempts to judge whether or not their agents' behaviours could fool players into thinking they were controlled by real people. These provided a pool of different studies which explored many games and agents. However, given the many differences between these studies, it is logically impossible to compare between them. This prompted further research with focus on the parameters themselves and even how the design and presented context of a game also affects the outcome. With player experience and game context, areas of interest in affect computing, deeply involved in these assessments, how are these two fields rarely connected? In this study, we explore this connection by providing a novel way of assessing human-like play using affective computing techniques. It accomplishes this through moment-to-moment assessment of participants' gameplay, showing that it can provide richer information. In addition, it demonstrates that this new method can not only compliment existing techniques, but also to extract consistent features with linearly predicting factors.**

1. Introduction

Artificial Intelligence (AI) remains one of the most sought-after subjects in computer science and with research presenting many different goals. Recent focus has remained within creating agents to compete with or aid players in different game tasks. As a result, many algorithms and techniques have been developed to this end [1]. One particular field within this subject is the development of agents that behave more human-like. However, when compared to the existing techniques, human-like behaviour still lacks the sufficient amount of academic focus. The possibilities for this could be the complexity behind the definition of "believable" and how to assess, parameterize and reason about it as well. Thankfully some research has been attempting to address these problems, by using the original Turing Test [2] and adapting it to assess how believable NPCs are [3], [4].

However, this is not sufficient as many of these previous studies display a wide diverse range of competitions, parameters and evaluation methods. Thus, they are not directly comparable and do not definitely tell us which one has the most believable bots nor provides a reasonably measure of how effective was the assessment approach. Another study, however, has shifted the focus to how parameters affect the outcome of the assessments, showing that these are important and do make a difference [5]. In addition, studies such as these show us that participants perceive believability differently to each other - each participant has a different internalized definition of what "believable" is. Another important field which focuses on how players perceive games and studies their experience is affective computing [6]. With the aim of understanding how humans play and to model their interactions with the environment based on their data, it seems only logical to pair these fields together.

The assessment of agents' believability is an important step and there is still no current accepted format. As such, this field needs to be explored further to find out what it is, which parameters are available, and why we should use them. Affective annotation can help us understand how players perceive believability within games which can only aid this quest. The already mentioned studies, and many others, ask questions such as "Is this believable?" about entire videos of gameplay. With this in mind, what are people reacting to? Were there moments that made participants judge bots as not believable? Does this mean any agent that is considered not human-like doesn't have any believable characteristics? What criteria makes them behave like a person would? There are many tools within affective computing that can help us towards answering these questions. One such tool is PAGAN [7]. A tool that allows researchers and participants to annotate moment-to-moment videos of gameplay, using the affect of choice.

For this research, we will be diving deeper into believability as a concept and how it has been assessed thus far. Then, a short summary of affective computing and one of its many tools: PAGAN. How this study combines both fields follows with the description of the game used for the assessments. Next, this paper discusses the data collected: including how it has been preprocessed and which gameplay features are present. It concludes with our findings and subsequent discussion for future work.

2. Background

2.1. Believability

When considering the concept of believability, it remains a broad notion by itself, lacking a generally accepted definition. Some of the earliest definitions date back to 1992 [8], in which believability is described as the "suspend disbelief" of a user. Other possibilities are further explored in [9] that range from broad definitions - where character believability is about providing illusion of life [10] - to more detailed definitions - where believability takes into account several elements such as emotions, personalities and intentions [11]. However, Loyall [12] attempts a different definition: believability is about supplying representations of personalities that are expected by the spectators.

When considering such definitions for video games, their complexity makes it harder to create a new definition. Researchers have attempted to provide us with said definitions [4], [13] and divided this concept into two classes: *Character Believability* and *Player Believability*. The first considers a fully autonomous agent/bot which has no human controlling it, but the agent acts in a believable way to a human observer. Player believability means that a character gives the illusion that a human player is controlling the agent, rather than the computer.

Given how complex this term is, evaluation becomes complicated albeit still important. To the author's knowledge there is still no accepted method of evaluating believability. There have been a few attempts at establishing this evaluation: based on generated criteria [14] and based on subjective assessment [13], [15]. These studies presented a diverse range of competitions, parameters for assessment and evaluation techniques. Thus, it becomes difficult to compare those bots and know which one is the most human-like. They have also focused purely on behaviour of characters. It seems common to believe that agent's believability is only dependant on the AI that controls it [16]. This sort of assumption has given way to two other contributions, ones that we shall focus more, that show that other factors, external [5] and internal [17] to the game, affect it as well. The first contribution [5] focused on how the parameters themselves can affect the accuracy of the assessment. The study was performed with a lower consideration for evaluating the human-like abilities of the bots and players and, instead, focus on how changing multiple factors – game target audience, camera perspective, player experience, length of videos, etc. – can change the outcome of the assessment. The second contribution [17], focused on how changing the agent's environment affects their believability – for both players and AI. For this study, the authors used a platform game – variant of Super Mario Bros – and asked participants to annotate the believability of the game's playthroughs which featured different level configurations – number of enemies, number of gaps, placement of both and many others. The goal of this study was to model player believability through machine learned representations of the obtained annotations.

Figure 1. Screenshot of RankTrace annotation

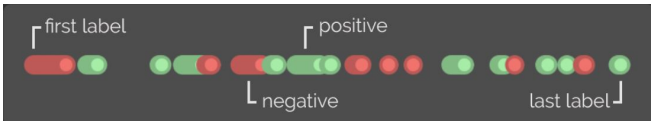

Figure 2. Screenshot of BTrace annotation

2.2. Affect Computing and Annotation

How do these techniques helps us understand how players perceive believability? With the complexity of this concept, if a player judges a bot as 'believable', what exactly made them think so? This brings us to the affective computing [6] and its possible role in answering those questions.

This is a field that concerns itself with understanding human affect and how to design interactions that are more aware of them. For this understanding, data is collected from the interactions between player and environment and labels provided from assessing these. It will then use the collected information to adapt people's experiences. With this in mind, one could 'label' believability and attempt to develop models from that data [5].

Tools and techniques are necessary to collect said information. The best way to do so would be through a quick but unobtrusive way. For this work, the choice of tool can be seen described in the following section.

2.2.1. PAGAN. PAGAN [7] is a free easy-to-use tool for multi-purpose video annotation. It does not require any installation, researcher presence or knowledge of any programming language. The researcher uploads the video(s) he/she wants annotations for; fills in a form with options for the study – this includes title, description, annotation type, project source, sound/no sound and many others – and share a link with the participants, once ready for access. Of the annotations available, our focus will be on RankTrace [18] and BTrace [19]. The participants will then access this link and start a session. The baseline needed to save the results is at least 25% - this means they must have seen at least 25% of the video and annotated it to log it – whenever and wherever they prefer. This provides easy access to any experiment for anyone in any part of the world and, most importantly, a way for generalising affective analysis. This has been previously tested for measuring moment-to-moment interest in 3 different videos. These displayed recorded gameplay sections of *Apex*, the season 8 trailer for *Game of Thrones* and a conversation between a human participant and "Spike" (a virtual agent from the *SEMAINE* database [20]). As for the annotations being used:

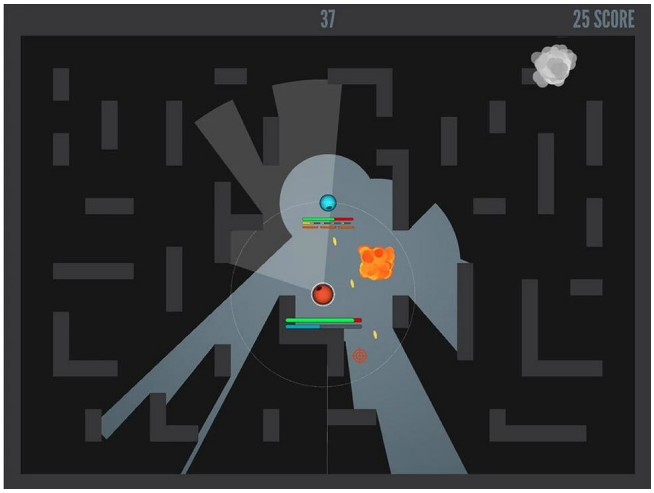

Figure 3. Game screenshot of *MAZING* from [21]

- RankTrace is related to ordinal affect annotation and its implementation is based on the work of Lopes *et al.* [18], which can be seen in Figure 1. Participants use the mouse wheel interface in order to control the positive and negative changes of affect during their subjective evaluation. For example, this could be to measure game frustration, arousal or any other type of player experience.
- BTrace, which stands for Binary Trace, is a simpler one-dimensional alternative. This tool is based on AffectRank [19] and can be seen in Figure 2. Here, participants use the 'up' and 'down' keys for positive $(+1)$ and negative (-1) changes respectively. This makes BTrace an alternative to relative annotation and, instead, a binary one.

3. Protocol

To the author's knowledge, moment-to-moment believability assessment has remained untested. Thus, the aim of this work, is to use an existing game and collect gameplay affective annotations. This is achieved by playing *MAZING* and annotating the opponent's believability in the recorded gameplay.

3.1. *MAZING* Game

MAZING[1] is a 2D top-down maze shooter [21]. In this game, the player can move freely around a maze and score points against an artificial agent by attacking it - see Figure 3. There are two different ways of damaging the opponent: by shooting projectiles at it or throwing bombs. The latter creates fires in the landing area which can continuously damage whoever stands on it for 5 seconds. Both attacks possess cooldowns. The player can also dash and its movement is faster than the agent's. There's also a partial fog-of-war which is illuminated by a player's field of view cone. To avoid losing score points, it must not pass through the fire carpets or collide with the monster. If there is a collision, both re-spawn at the original starting locations.

As for the agent, it starts off searching for the player by wandering around the maze. Once it picks a spot in the map, it travels there using a basic search behaviour. If the player comes into the agent's field of view or auditory range, it chases it. The agent doesn't shoot or throw bombs, it possesses only movement and low-level decision making. It also avoids player's shots/bombs and has many hit points before respawning.

One important feature that drives the agent's behaviour is its frustration model. The frustration model being used in this study is the one introduced in [21]. It was implemented to provide human-like characteristics to the agent and, depending on the "frustration score", the agent's previous features are affected. Most notably, the enemy's field of view will decrease - given the increased 'focus' when searching for the player - but the likelihood of 'hearing' the player increases. The agent will also increase its movement and rotation speed, and risk going through more dangerous paths to get to the player. Frustration levels increase when the agent finds the player and is unable to reach it throughout the game. However, it decreases when it loses the player and returns to its searching behaviour. For full details, see [21]

3.2. Study Procedure

In a novel study, the decision was made to explore a moment-to-moment assessment of believability - given that previous assessments involved overall judgement on single videos [5], [11], [15].

To achieve this, the study was set up using PAGAN's framework, proving its generality and usefulness within affective research. Participants were given one of two links: one for Btrace annotation and another for RankTrace annotation. After the instructions, the participants get to play a game tutorial first - to get acquainted with it - and 2 play-annotation sections follow. Each play lasts 1 minute, with the video being recorded during this time. The opponent is also different in each session, with one session having an agent with no frustration score and the other with an agent with 50-100 frustration score. The order of the agents is randomized - participants could play against the 'frustrated' agent on the first session or the second.

The recorded gameplay is then presented to the participant to annotate. During the video-labelling task they are asked to label their opponent's gameplay in terms of believability. The term 'believable' meaning their opponent is playing like a human would in the given situation. Thus, allowing the participants to label the change of believability.Once the experiment is finished, the participants will be asked for their preference over the 2 videos and an exit survey is presented to them.

1. https://davidmelhart.com/projects/mazing/index.html

4. The Data

In this section we describe the dataset, how preprocessing was done to clean the data and which features were collected from the gameplay. This data is made up of information on both the player and the bot activity throughout the session. It was collected over a period of 2 weeks during November and December 2020. The raw data is made up of a total of 86 participants - 36 for BTrace and 50 for RankTrace. The amount of datapoints varies between participants, depending on how many annotations were made per session. There were $13,640$ datapoints for BTrace and $15,884$ for RankTrace.

4.1. Preprocessing

To prevent bias in our data, the dataset had to be cleaned of invalid datapoints. The first validation involves deleting the datapoints in reloaded sessions. This happens when participants refresh the webpage - when they get stuck for example - and start annotating the video from the beginning again. The datapoints discarded are the ones up to the reload moment. A total of 11 entries were discarded. After this the datapoints where the sessions were too short or had unresponsive participants - made less than 10 annotations - were also dropped. Those were $2 + 7$ sessions for BTrace and $7 + 9$ for RankTrace. A further 10 BTrace sessions and 16 RankTrace sessions were dropped because they had nonexistent matching annotations.

Given these annotation methods and following previous work on processing this type of data [18], the gameplay is split into several time windows. It is from these windows that we extract statistical features. Once the time window was picked - 3000ms - we applied 3 different annotation metrics: the mean value, the amplitude (max-min difference) and average gradient. For example: divide a session into 3 second windows and average the annotation result within that window. The first allows us absolute extraction and the second and third allows us a relative change measurement. A 1 second delay was also applied between moment and annotation value. Finally, the features and the believability values were normalized.

The clean, re-sampled and final dataset has 23 participants, 42 sessions and 840 datapoints for BTrace. It also has 27 participants, 40 sessions and 801 datapoints for RankTrace.

4.2. Extracted Features

The game extracts 54 features during gameplay. This allows us to measure player behaviour, opponent behaviour and context. Thus, they range from bot actions - such as the distance it travelled - to the player's - such as positioning. A full list of these characteristics can be seen in Table 1.

The names attributed to the features were intended to be as straightforward as possible. However, for the ones that might not be as simple to understand: the features which finish with 'OnCD' signify 'On Cooldown' - this means the player attempted these actions while they were not available yet; 'inputIntensity' stands for the number of times keyboard keys and mouse have been pressed; and 'inputDiversity' stands for the keys that have been pressed so far.

4.3. Results

To calculate the correlations between the features and the believability annotation values we use Kendall's τ correlation coefficient with p-value< 0.05. Despite the believability assessment being done on the opponent, all features are measured since context is important: the situation the player is in also influences the outcome of the assessment.

Each feature was also checked for each of the 3 feature annotation metrics - mean, amplitude and gradient. The values were also done globally - on all data available. Overall, the mean has provided more significant results than amplitudes and gradients. These results were the same for both BTrace and RankTrace, with BTrace having 24 significant results when using the mean, 7 when using amplitudes and 2 when using gradients. RankTrace showed 17 significant results when using the mean, 14 for amplitudes and 11 with gradients.

There's also consideration for the value of the correlation coefficient of Kendall's to explore how strong the relationship is between the features and their believability. The interest being on both positive (linear) and negative (inverse relationship). For simplicity, the top significant results - with highest positive and highest negative - are displayed in Tables 2, 3 and 4. These are also displayed side-by-side for comparative measures. In bold are the features which are consistent between both RankTrace annotations and BTrace annotations in the top places.

4.4. Discussion

To explore if there were any correlations between the extracted features and the assigned believability values by the participants, Kendall's τ was chosen as the test for statistical dependence.

Taking the previous study on RankTrace for data handling [18], some of the same techniques were applied to BTrace. Despite the fact that these were only applied to RankTrace before, applying them to BTrace allows us to draw some comparisons between them - since they go through the exact same pre-processing. As seen, Table 2 presents the most consistencies between both tools with several features on top values. The relationships show that there are some predicting factors of believability to this data. The consistent features are mostly bot related - which is expected since we requested participants to annotate its behaviour - with only one being player related - the score. This seems to show that the most believable aspects of the bot correspond to when it is engaged in chasing the player. That is understandable since the more it sees and chases the player, the more interactions and focus behaviour it displays. This is in contrast to when it doesn't show as much activity: note the 'idleTime' shows an inverse relationship and,

TABLE 1. FEATURES EXTRACTED FROM MAZING

Bot		Player		General and Others
botLostPlayer	botDistanceTraveled	playerDistanceTravelled	playerBurning	cursorDistanceFromPlayer
botSpottedPlayer	botPositionX	playerPositionX	playerHealing	cursorDistanceFromBot
botBurning	botPositionY	playerPositionY	playerDeltaHealth	cursorDistanceTraveled
botDeltaHealth	botRotation	playerRotation	playerDied	cursorPositionX
botDied	botSpeed	playerHealth	shotsFired	cursorPositionY
botHealth	botRotationSpeed	playerIsDashing	bombDropped	onScreenFires
botFrustration	botViewAngle	playerTriesDashOnCD	gunReloading	onScreenBullets
botRiskTakingFactor	botViewRadius	dashPressed	bombReloading	timePassed
botTakingRiskyPath	botSearching	playerTriesToFireOnCD	playerTriesToBombOnCD	score
botSeeingPlayer	botSearchTurns			inputIntensity
botChasingPlayer	botHearingRadius			inputDiversity
botDistanceFromPlayer	botHearingProbability			idleTime

TABLE 2. TOP (POSITIVE AND NEGATIVE) KENDALL τ RESULTS WITH SIGNIFICANCE (p-VALUE< 0.05) FOR MEAN ANNOTATIONS FOR BOTH RANKTRACE AND BTRACE. CONSISTENT FEATURES IN BOLD.

Features RankTrace (Mean)	kendallτ	p-values	Features BTrace (Mean)	kendallτ	p-values
botChasingPlayer	0.15777	0	**botChasingPlayer**	0.25579	0
score	0.11119	0	**botSeeingPlayer**	0.23539	0
botSeeingPlayer	0.10056	0.00012	**botDistanceTraveled**	0.20168	0
botDistanceTraveled	0.0986	4,00E-05	inputIntensity	0.16454	0
botPositionX	0.0937	9,00E-05	**score**	0.14725	0
timePassed	0.09326	0.0001	inputDiversity	0.13952	0
cursorPositionX	0.06844	0.00432	playerDistanceTravelled	0.13675	0
botPositionY	0.06654	0.00551	botHearingProbability	0.12067	3,00E-05
cursorDistanceTraveled	0.05938	0.01327	**timePassed**	0.11434	2,00E-05
playerTriesDashOnCD	0.05774	0.04669	botRiskTakingFactor	0.10243	0.00014
botSpeed	-0.0479	0.0475	cursorDistanceFromPlayer	-0.07521	0.00459
idleTime	-0.06103	0.01491	playerHealth	-0.11073	0.00025
botHealth	-0.0814	0.00089	**botHealth**	-0.12298	1,00E-05
botRotationSpeed	-0.08327	0.00057	**botDistanceFromPlayer**	-0.14651	0
botDistanceFromPlayer	-0.10549	1,00E-05	**idleTime**	-0.16634	0

TABLE 3. TOP (POSITIVE AND NEGATIVE) KENDALL τ RESULTS WITH SIGNIFICANCE (p-VALUE< 0.05) FOR AMPLITUDE ANNOTATIONS FOR BOTH RANKTRACE AND BTRACE. CONSISTENT FEATURES IN BOLD.

Features RankTrace (Amp)	kendallτ	p-values	Features BTrace (Amp)	kendallτ	p-values
botFrustration	0.10235	0.0004	shotsFired	0.09354	0.00173
botBurning	0.08734	0.00354	**bombReloading**	0.07727	0.02253
onScreenFires	0.08551	0.0021	onScreenBullets	0.07122	0.01645
playerDeltaHealth	0.08507	0.00387			
playerDied	0.08507	0.00387			
botLostPlayer	0.08507	0.00387			
botSpottedPlayer	0.08507	0.00387			
bombReloading	0.08475	0.00695			
botHearingProbability	0.08397	0.00349			
botTakingRiskyPath	0.08042	0.01144			
botHealth	-0.05317	0.04708			
playerDistanceTravelled	-0.05438	0.03802	timePassed	-0.0708	0.01188
botHearingRadius	-0.05828	0.02738	botFrustration	-0.07267	0.01758
botViewAngle	-0.05959	0.02416	botHearingProbability	-0.08608	0.00476
cursorDistanceTraveled	-0.06126	0.01945	botSearchTurns	-0.09798	0.00166

so does 'botHealth' and 'botDistanceFromPlayer'). Future work will explore what role this labelled data can have in modelling and predicting believability in games. In addition, this data was processed on global results. It should also be explored how it would affect the outcome if it was processed on a per session/per participant basis. Would it show the same consistent results? Or would it make more player's experience specific?

BTrace showed not only cleaner data (pre-processing stage) but it is also presenting more correlations overall on Table 2 (after processing). This tool was developed to make the annotation process simpler [19], which may derive in a more intuitive usage. Another study could investigate players participating in both BTrace and RankTrace annotation and answer that same question. However, when comparing the number of correlations between BTrace's tables, it has substantially less on Table 3 and 4. These also have less correlations when compared to their RankTrace counterpart, with only one consistent feature in Table 3. In addition,it is also worth exploring other ways of processing BTrace's time

TABLE 4. TOP (POSITIVE AND NEGATIVE) KENDALL τ RESULTS WITH SIGNIFICANCE (p-VALUE< 0.05) FOR AVERAGE GRADIENT ANNOTATIONS FOR BOTH RANKTRACE AND BTRACE. NO CONSISTENT FEATURES.

Features RankTrace (Grad)	**kendallτ**	**p-values**	**Features BTrace (Grad)**	**kendallτ**	**p-values**
botSeeingPlayer	0.06884	0.00799	botRotationSpeed	0.05786	0.02287
playerDistanceTravelled	0.06636	0.00533			
botRotation	0.05625	0.01815			
inputDiversity	0.05387	0.02382			
cursorDistanceTraveled	0.05103	0.03214			
inputIntensity	0.04701	0.0484			
botViewRadius	-0.04771	0.04833			
idleTime	-0.06001	0.01593			
botSpeed	-0.06077	0.01135			
playerHealth	-0.06409	0.02111			
onScreenFires	-0.07125	0.00478	botFrustration	-0.05873	0.03298

windows.

Finally, with the collection of preference over both opponents that each participant faced, the next step is to investigate which way is better for assessing believability. Previous methods (preference based included) have focused on the whole gameplay. This method provides a moment-to-moment assessment of believability, taking into account many features in the process. The data seems to be richer, but another possibility would be to compute an 'overall' believability score per video and check if it matches the preferred opponent. It could simply not be comparable and both provide different information. This technique can, however, compliment state-of-the-art techniques for even richer information.

5. Conclusion

This paper discusses the state-of-the-art believability assessment techniques and the need for more reliable and comparable methods. For that, it suggests the use of PAGAN: a straightforward tool with different annotation techniques that can be integrated with the previous believability assessment approaches. This would allow for the novelty of collecting annotations on moment-to-moment gameplay videos. To prove this concept, our study asks participants to play a simple game and annotate how believable their opponent is. It shows that PAGAN can be used widely within affective analysis and, given the consistency of the data, it opens the possibility of using these annotations to predict believability and model it. With believable behaviour being a sought-after feature both within video games and research, this would impact both fields positively. Overall, this work can not only support existing research but it also has the opportunity to expand it.

Acknowledgments

This research is supported by the IEEE CIS Graduate Student Research Grants and the EP/L015846/1 for the Centre for Doctoral Training in Intelligent Games and Game Intelligence(IGGI) from the UK Engineering and Physical Sciences Research Council (EPSRC).

References

[1] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[2] A. M. Turing, "Computing Machinery and Intelligence," in *Parsing the Turing Test*. Springer, 2009, pp. 23–65.

[3] P. Hingston, "A Turing Test for Computer Game Bots," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 1, no. 3, pp. 169–186, 2009.

[4] J. Togelius, G. N. Yannakakis, S. Karakovskiy, and N. Shaker, "Assessing believability," in *Believable bots*. Springer, 2013, pp. 215–230.

[5] C. Pacheco, L. Tokarchuk, and D. Pérez-Liébana, "Studying believability assessment in racing games," in *Proceedings of the 13th international conference on the foundations of digital games*, 2018, pp. 1–10.

[6] R. W. Picard, *Affective computing*. MIT press, 2000.

[7] D. Melhart, A. Liapis, and G. N. Yannakakis, "Pagan: Video affect annotation made easy," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 130–136.

[8] J. Bates, *The nature of characters in interactive worlds and the Oz project*. School of Computer Science, Carnegie Mellon University Pittsburgh, PA, 1992.

[9] F. Tencé, C. Buche, P. De Loor, and O. Marc, "The challenge of believability in video games: Definitions, agents models and imitation learning," *arXiv preprint arXiv:1009.0451*, 2010.

[10] F. Thomas, O. Johnston, and F. Thomas, *The illusion of life: Disney animation*. Hyperion New York, 1995.

[11] M. O. Riedl and R. M. Young, "An objective character believability evaluation procedure for multi-agent story generation systems," in *International Workshop on Intelligent Virtual Agents*. Springer, 2005, pp. 278–291.

[12] A. B. Loyall, "Believable agents: Building interactive personalities." CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, Tech. Rep., 1997.

[13] D. Livingstone, "Turing's test and believable ai in games," *Computers in Entertainment (CIE)*, vol. 4, no. 1, pp. 6–es, 2006.

[14] H. Warpefelt, "Evaluating the believability of agents in virtual worlds." ACAI, 2013.

[15] J. Togelius, G. Yannakakis, S. Karakovskiy, and N. Shaker, "Believable bots: Can computers play like people," 2012.

[16] J. D. Miles and R. Tashakkori, "Improving the believability of non-player characters in simulations," in *Proceedings of the 2nd Conference on Artificiel General Intelligence (2009)*. Atlantis Press, 2009.

[17] E. Camilleri, G. N. Yannakakis, and A. Dingli, "Platformer level design for player believability," in *2016 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2016, pp. 1–8.

[18] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 158–163.

[19] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2015, pp. 574–580.

[20] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011.

[21] D. Melhart, G. N. Yannakakis, and A. Liapis, "I feel i feel you: A theory of mind experiment in games," *KI-Künstliche Intelligenz*, vol. 34, no. 1, pp. 45–55, 2020.