Discrete versus Ordinal Time-Continuous Believability Assessment

Cristiana Pacheco*, David Melhart†, Antonios Liapis†, Georgios N. Yannakakis† and Diego Perez-Liebana*
* Queen Mary University of London, London, UK
Email: {c.pacheco, diego.perez}@qmul.ac.uk
† Institute of Digital Games, University of Malta, Msida, Malta
Email: {david.melhart, antonios.liapis, georgios.yannakakis}@um.edu.mt

***Abstract*—What is believability? And how do we assess it? These questions remain a challenge in human-computer interaction and games research. When assessing the believability of agents, researchers opt for an overall view of believability reminiscent of the Turing test. Current evaluation approaches have proven to be diverse and, thus, have yet to establish a framework. In this paper, we propose treating believability as a time-continuous phenomenon. We have conducted a study in which participants play a one-versus-one shooter game and annotate the character's believability. They face two different opponents which present different behaviours. In this novel process, these annotations are done moment-to-moment using two different annotation schemes: BTrace and RankTrace. This is followed by the user's believability preference between the two playthroughs, effectively allowing us to compare the two annotation tools and time-continuous assessment with discrete assessment. Results suggest that a binary annotation tool could be more intuitive to use than its continuous counterpart and provides more information on context. We conclude that this method may offer a necessary addition to current assessment techniques.**

***Index Terms*—Believability, Turing Test, Human-Like Agents, Time-Continuous Annotation, Digital Games, Assessment**

I. Introduction

Human-like behaviour is a crucial property of artificially intelligent agents in the field of Human-Computer Interaction, as these agents can aid automated testing [1], provide a challenging competitive partner [2], collaborate with users [3] or even support a therapy process [4]. However, developing these agents remains a central challenge in the field of artificial intelligence (AI). A major roadblock is the lack of a universally accepted definition of *believability*. Most studies use behavioural cues to analyse the agent's perceived decision-making capacity and expression of believability [5].

Game research is a major frontier in the study of believability as games provide a rich testing ground for emergent interactions. In the field of game research, studies have attempted to establish core concepts and methods for assessing believability [5]–[7]. The most common method has been the adaptation of the Turing Test [8], and the separation of agent believability into two concepts: player (or user) believability and character (or non-player) believability [5]. The former concept refers to the perception that an artificial agent is human-controlled, even when it is autonomous, whereas the latter refers to recognising human-like *behaviour* in autonomous agents. Even though they follow a similar methodology, these studies present a wide range of approaches to assessing believability—e.g. binary decision, rating, ranking, forced preference choice, both when participants are observing and playing the games [5], [9]. While most of these studies focus on the agent and game genre, the environment is often overlooked [10]. Another major limitation of these studies is their reliance on a discrete representation of believability that rarely takes context into account [5], [11].

To address the aforementioned limitations, we propose a method of time-continuous annotation of believability assessment relying on affective computing techniques. In this study, we focus on the less explored *character believability* [9]. We test our method on an asymmetric top-down shooter game called *MAZING* [12]. Even though the visuals of the game are abstract, the agent was designed to exhibit human-like behaviour. This setup provides a testbed with a good amount of complexity—as suggested by the literature [5]—while simplifying the problem by focusing on the game context and agent behaviour. A user study was designed where participants are asked to play this game and assess their opponent's moment-to-moment believability, followed by a questionnaire with a discrete evaluation method. For the labelling task, participants are randomly assigned to two different tools, using the *Platform for Audiovisual General-purpose ANnotation* (PAGAN) annotation framework [13]: *BTrace*, a binary tool based on *AffectRank* [14], and *RankTrace*, a continuous unbounded tool designed to collect time-continuous ordinal data [15]. The questionnaire features a subjective method of preference between videos [5]. This permitted a comparison between classical methodologies and time-continuous tools included in PAGAN. Our analysis shows a correlation between continuous and discrete believability assessment, with BTrace leading the results, reinforcing conclusions drawn by our complementary study [16]. The novelty of this work is the introduction of time-continuous assessment for human-like agents. Our results suggest that the methods presented here are a worthy addition to the existing assessment techniques and a step towards a normative protocol of time-continuous believability evaluation,

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).

978-1-6654-0021-3/21/$31.00

with binary time-continuous annotation seeming a more suitable choice for the context of believability assessment.

II. Background

This section discusses relevant research on believability, emotional behaviour simulations and assessment protocols.

A. Defining Believability

While there were several attempts to create a unified definition, the term *believability* still lacks a precise description. Some provide more high-level definitions, defining believability as an illusion of life and a *suspension of disbelief* that is dependent on viewers' expectations [17], [18]. Others opt for more specific interpretations, where intentionality and rationality [19] or a balance between predictability, randomness, and behaviour exaggeration [20] are seen as core components for believability. Additionally to definitions focusing on behaviour, some studies have also highlighted the effect of the environmental context on the agent's believability [20], [21].

In digital games, research of believability is often focused on the study of Non-Player Characters (NPCs) and their interactions with players [22]. Commonly observed characteristics of believable agents are explicit intention, consistent behaviour, and emotional expression [19]–[22]. Lankoski and Björk extend this list further with "natural language" [22]. However, this criterion is debatable, given that not all NPCs require speech capacity to elicit an emotional response. Moreover, it has been suggested by previous literature that research into believability would benefit from decoupling narrative and aesthetics from gameplay [5], [23]. This way, the behavioural component of believability could be studied separately from aesthetic components, such as a human-like appearance, eliminating some of the fuzziness of the concept.

B. Simulating Emotional Behaviour

To create more adaptable and socially aware AI, we need agents that implement robust models for human-like emotion regulation and manifestation [21]. Research in the field has been focusing primarily on adapting *Appraisal Theory* [24], [25] to this end. Appraisal Theory focuses on emotions as functions of an evaluation process of antecedent events. One of the most popular frameworks of emotional appraisal is the *Ortony–Clore–Collins* (OCC) model [26], which describes the strength of emotions as a function of actions, consequences, and the environmental and social context. Several computational frameworks adapted OCC to create models for artificial emotion regulation. Examples of these include *EMotion and Adaptation* (EMA) and *Fearnot AffecTIve Mind Architecture* (FAtiMA). The former uses both the manifestation of emotion and its influence on future actions to model appraisal as a uniform, but temporally causal process [27]. The latter, by contrast, uses a two-tier system: one based on instant events and another on the chance of future success [24].

Less focus has been given to *simulating* behavioural manifestations of emotion. To address this research gap, Melhart *et al.* investigated the recognition of emotional behaviour of artificial agents [12]. In their experiment, an agent was designed to exhibit frustration based on the theory of *Computer Frustration* [28]. Their game featured exaggerated behaviour and many visual cues to help recognise emotion without the use of a human-like appearance [20]. Participants were asked to annotate the perceived frustration level of the agent. The results of their experiments highlight the importance of context in the appraisal of NPC behaviour.

C. Assessing Believability

Perhaps the earliest and most popular test developed for human-like evaluation in a machine is the Turing Test [8]: a test where a participant communicates with another person and a computer through text and deems the bot intelligent if the participant is incapable of distinguishing which one is which. Subsequently, researchers disagreed that this was the correct way of spotting intelligence [29], [30] and it remained criticised years later, especially within digital games, as it was far too simple for such complex environments [31].

Several attempts have been made to establish an assessment protocol for believability. Some have been based on formal criteria [7], [32], such as how the agent navigates the world, and how it reacts to environmental or social changes. Other attempts, and perhaps the most common, are based on subjective assessment [23], [31] which involves the observation of an agent and filling a questionnaire. In general, these studies were using the Turing Test adapted to gameplay instead of text-based communication. Perhaps character believability receives less attention due to the required level of realism [21], since participants know the agents are artificial and are usually inquired about their behaviour [9], [33], [34]. While it is a common assumption to attribute believability primarily to agent behaviour [35], more recent studies have indicated that the context of the interaction has just as much effect on the perception of believability [10], [11]. For example, Camilleri *et al.* showed that the agent's believability is highly dependent on the design of the level—e.g. the number of enemies, gaps, entity placement, etc. [11]. In addition, Pacheco *et al.* showed that changing how the test itself is presented (camera perspective, player experience, length of videos, etc.) changes the outcome of the assessment [10].

While assessment is discrete in believability research, in affective computing time-continuous annotation has become more and more popular over the years [36]. Time-continuous annotation protocols capture moment-to-moment changes in participants' affective state using a wide array of labelling methods. While discrete techniques are still used to reduce noise and to achieve a higher score consensus [14], time-continuous methods tell a more nuanced story about the data as they capture the moment-to-moment variations of participants' experience [36]. Even though popular techniques are based on the classical Likert-scale [37], a usability study of Melhart *et al.* showed that unbounded ordinal annotation [15] is more intuitive, and a binary labelling strategy leads to higher inter-rater agreement [13]. In this study, we rely on these type of annotation methods that are implemented in the PAGAN

framework as RankTrace (unbounded) and BTrace (binary). Although believability assessment is not a traditional affective computing task, Hamdy *et al.* suggests a connection between believability, emotion, and behaviour towards an overarching architecture [21]. In this paper, we rely on this connection to investigate believability under the lens of affective computing. We combine the time-continuous annotation techniques seen in affective computing [13] with the discrete techniques seen in believability assessment [5].

III. Study Protocol

The approaches mentioned in section II-C present adaptations of the Turing Test for agent evaluation. Despite the necessary contributions to the field, it presents its own challenges. The first problem lies in the terms 'believable' and 'human-like'—as these remain vague—and a participant's own perceptions of believability which allow bias in their evaluation [5], [10]. In addition, a lack of an accepted framework remains given the range of options provided and the lack of comparable results [9], [10]. Lastly, these studies provide only an overall view of believability based on entire sessions [5], [9], [10], thus lacking a lower-level understanding of the agent in the context given.

Our suggestion is to address these problems by providing an additional technique: moment-to-moment believability annotation. This allows us to tackle the ambiguity of the original process, build towards established protocols and compare with future results. Given the novelty of this process we will follow the suggestions of previous literature [5], [23] and use a single-player game without narrative or natural language. This section presents the game, the experimental protocol as well as the methods for collecting believability annotations and for cleaning the data.

A. Environment: MAZING

MAZING is an asymmetrical top-down shooter where the player is chased through a maze by an AI controlled opponent. The player scores points when damaging and killing the AI opponent, which can be done with bullets or bombs. Bombs, as seen in Figure 1, are slower projectiles that explode and leave a fire in the area for a short period of time. Both player and opponent can take damage when walking over it. The opponent's goal is to chase the player. The opponent kills the player if the two collide; if this happens, the game resets. The agent does not possess the same skills as the player: it moves faster, it has no weapons and it has two sensory systems (a narrow field-of-vision and an auditory system). Its behaviour depends on the situation: if the player is out of sight, it moves randomly through the level seeking the player; if the player is in sight it chases them using the closest path. However, if the closest path is through a fire, it follows it if its health is high or if the alternative is a much longer path. In addition, an abstract model of Computer Frustration [28] influences its sensors and decisions. A more frustrated agent will have a slimmer field of view but more precise audition; it will take more risks and move more erratically. In the user study described in this paper, two different AI opponents were used, with different levels of frustration; the order of opponents was randomized across participants.

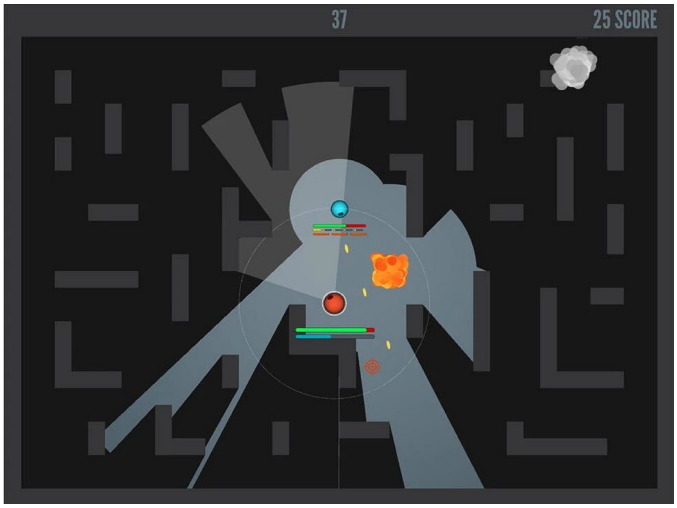

Fig. 1. MAZING Screenshot. The blue dot is the player, the red is the computer agent. The light grey area is the player's field-of-vision. Both player's attack modes are visible (yellow projectiles and a fire) in the middle and an extinguishing fire in the top right corner.

B. Data Collection and Preprocessing

The user study is conducted over two rounds of gameplay and subsequent annotation. Before the user study begins, participants are allowed to learn and test the controls of *MAZING* for as long as they wished. For the study, participants play a 1-minute round of the game against an opponent. Once their game session is completed, then a video replay of it is shown to them for moment-to-moment annotations of the opponent's believability. Participants are instructed that "believability means your opponent is playing like a human would in the given situation". This process is repeated a second time with an opponent exhibiting a different gameplay behaviour—randomly assigning one of two frustration levels (see Section III-A). Once the experiment is finished, the participants are presented with an exit survey asking for their preference over the two videos in terms of believability—the choices being *First game*, *Second game*, *Both were equally believable* and *Neither were believable* and some demographic questions.

This process was carried out through PAGAN [13], a tool for online affect annotation. This framework has been modified to allow the whole experiment to take place in PAGAN via a single link. This study tested two time-continuous annotation methods: a binary discrete annotation tool called *BTrace*, or an unbounded continuous annotation tool named *RankTrace* [13] (see Fig. 2). Participants were assigned randomly between the two, and completed the whole experiment using only one annotation method. The data collected consists of telemetry (i.e. game data such as the player's score, the opponent's health or the buttons pressed), the believability values for BTrace (binary) or RankTrace (continuous), the overall believability

(a) RankTrace annotation

(b) BTrace annotation

Fig. 2. PAGAN data collection interfaces used for the experiments.

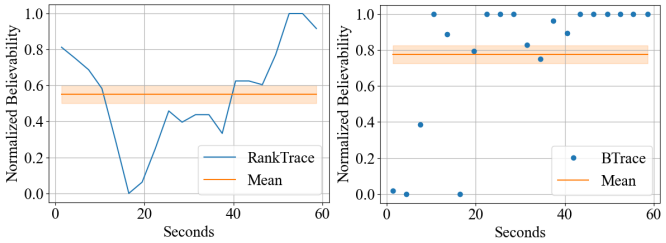

Fig. 3. Normalized traces of believability produced by RankTrace (left) and BTrace (right). Area of uncertainty is in orange.

preference between the two videos and demographics. The telemetry data could shed light on which player or agent properties affect believability; as this study focuses on the quality of the annotation traces themselves, we do not use the telemetry data in this paper.

Because PAGAN only records changes in the annotation trace due to the limitations of the online format, the annotation traces are first resampled at a rate of 250ms. During the resampling empty values of RankTrace are forward filled and empty values of BTrace are filled with zeros to respectively preserve the continuous and discrete nature of the annotation traces. Because 250ms time-windows are too short to meaningfully capture changes in the perceived beliveability, we further aggregate the annotation traces into 3-second time-windows. We chose the window size based on convention set by previous papers using RankTrace [15], [38], [39] and the MAZING testbed game [12]. The time-windows are calculated consecutively, based on the mean value of those windows. The annotation values of RankTrace—which are unbounded—are also normalised to [0,1] via min-max normalization on a per-video basis.

In order to clean the dataset of outliers, we use the *Dynamic Time Warping* (DTW) distance based on Makantasis *et al.* [40]. DTW is an algorithm that measures the similarity between temporal sequences. We use DTW to detect outliers in two steps. Initially we measure the distance to an artificial *inactive baseline*—all annotations at zero. We discard sessions that fall more than two standard deviations towards zero from the mean distance of the dataset to the inactive baseline. The cumulative DTW distance is then calculated for each session by summing up the DTW distance between the session compared to all other sessions. We discard those sessions that fall more than two standard deviations away from the average cumulative DTW distance of all sessions. This process removes unusual believability annotations with insufficient data that deviates from the annotators' consensus. Finally, participants that do not have two valid sessions and responses to the final questionnaire are also removed as both components are needed for comparison.

C. Matching Believability Traces with Believability Preference

To compare between the time-continuous data and the believability preference between sessions, we discretise the former. The data is discretised into two types: *High* believability and *Low* believability. To calculate the number of high and low instances per session, we first calculate the mean (μ) of a session's (normalized) annotations. We then apply an uncertainty threshold (ϵ). Annotations that fall within $[\mu - \epsilon,\ \mu + \epsilon]$ are treated as ambiguous (neither high nor low) and discarded. In Figure 3, we can see an example of both RankTrace and BTrace session annotations in blue. The orange section shows the mean with 2ϵ value range applied. The number of 3-second time windows above the shaded orange section are counted as *High* (H) and those below it as *Low* (L). A third metric, *Difference* (D), is calculated as $D = H - L$ and can be positive (when there are more instances of high believability), negative (when there are more instances of low believability) and zero (when high and low instances are equal).

IV. Results

A web link of this study was sent through the author's contacts following a purposive sampling approach. While 89 users interacted with MAZING, only 45 participants completed both game sessions and the final questionnaire. After the cleanup process described in Section III-B was complete, the remaining participants were 27; 11 for the RankTrace version and 16 for the BTrace version. Among the final 27 participants, most identified as male (78%) and 22% identified as female. The average age of participants was 29 years old and most were 20-29 years old (59%); participants' age ranged from 18 to 49. All participants had experience playing video games, with 48% playing games everyday, 33% playing games a few times a week and 19% playing games a few times a month.

A. Correlation Analysis

We treat the binary choice of believability preference between the first or the second session as the ground truth. We assign 1 if the first session is preferred, -1 if the second session is preferred, and 0 if no clear preference is given ("both" or "neither" options). We then calculate the correlation of binary preference with the differences between the metrics of believability traces described in Section III-C, taking the metric of the first session and subtracting the metric of the second session. We presume that a higher correlation with the ground truth points to a more accurate and meaningful time-

TABLE I
PEARSON'S CORRELATIONS BETWEEN DISCRETE BELIEVABILITY ASSESSMENT AND DISCRETISED CONTINUOUS ANNOTATIONS FOR BOTH TOOLS WITH DIFFERENT THRESHOLDS (ϵ). BOLD VALUES ARE SIGNIFICANT CORRELATIONS AT $\alpha = 0.05$.

ϵ	**Tool**	$H_1 - H_2$	$L_1 - L_2$	$D_1 - D_2$
0.00	RankTrace	-0.112	-0.323	0.073
	BTrace	**0.739**	**-0.568**	**0.694**
0.05	RankTrace	-0.112	-0.323	0.073
	BTrace	**0.739**	**-0.568**	**0.694**
0.10	RankTrace	-0.112	-0.323	0.073
	BTrace	0.282	**-0.568**	0.444
0.15	RankTrace	-0.112	-0.323	0.073
	BTrace	-0.187	**-0.568**	0.085
0.20	RankTrace	-0.112	-0.323	0.073
	BTrace	-0.092	-0.442	0.164
0.25	RankTrace	-0.112	-0.098	-0.033
	BTrace	-0.247	-0.17	-0.157

continuous annotation tool. The significance is reported via Pearson's product-moment correlation coefficient at $\alpha = 0.05$.

The uncertainty threshold (ϵ) presented in Section III-C can provide a cleaner dataset in case of ambiguous data points which are very close to the mean, but can also reduce the dataset significantly. We thus explore a broad variety of ϵ values and how they impact the correlation values of the two time-continuous annotation tools. Table I shows the results of the analysis, highlighting significant correlations. It is evident that BTrace has stronger correlations with the ground truth than RankTrace. In fact, RankTrace's highest absolute correlation remains near the lowest absolute correlations from BTrace. The impact of the uncertainty threshold is also obvious, especially for BTrace, as an over-aggressive threshold is likely to remove a large part of the dataset (omitting, essentially, all high values if the mean value is also high) and causing large fluctuations to correlation. The fact that high ϵ values cause many high values to be removed from the dataset is evidenced by the fact that difference in low values ($L_1 - L_2$) is unaffected at $\epsilon = 0.1$ and $\epsilon = 0.15$. For BTrace, the high positive correlations of the believability preference with high values and high negative correlation with low values is not surprising. The fact that this does not hold for RankTrace is more surprising, with $H_1 - H_2$ having a negative correlation to believability preference. The aggregated D values (measuring the difference between high and low instances within the trace) seems also to be robust for BTrace, attaining a higher absolute correlation value than low values and comparable correlation to high values; this means that we can use the summary metric D instead of having to observe both H and L.

B. Additional Analysis

The analysis of Table I showed that BTrace metrics as instances of high or low believability were highly correlated with the binary believability preference for low ϵ thresholds but deteriorated at higher thresholds. This indicates that there is substantial data loss at high uncertainty thresholds, likely due to higher mean values of the trace. As observed in Fig. 3, users' binary annotations tended to be more positive than negative which led to a higher mean value for the entire trace and thus high thresholds might extend past the upper limit (1) and remove most data points. To assess the impact of μ on the metrics of high, low and difference for the two tools, we ran another correlation analysis where the absolute midpoint of the value range (0.5) was used as the interim value from where the uncertainty bound was formed as $[0.5-\epsilon, 0.5+\epsilon]$. Correlations of believability preference with the metrics as calculated with the "neutral" uncertainty bound are identical for RankTrace as those at respective ϵ thresholds in Table I; therefore the μ for RankTrace does not impact the quality of results. For BTrace, correlations at all inspected ϵ thresholds investigated (from 0 to 0.25 at increments of 0.05) are the same as the respective correlations of Table I only at $\epsilon = 0$. This validates our assumption that the high μ values of BTrace led to the drop in correlations shown in Table I.

As a final experiment, we investigated the correlation between the binary believability preference and the difference in mean values of each gameplay trace (i.e. $\mu_1 - \mu_2$) for BTrace and RankTrace. The correlations showed similar trends as with all other metrics explored, with BTrace difference between means aligning better with the participants' preference with a significant positive correlation ($\rho = 0.667$) while for RankTrace there was a negative correlation ($\rho = -0.321$) below the significance threshold. Intuitively, more believable agents would have a higher mean trace value, although the fact that RankTrace users annotate in an unbounded fashion means that this metric is less reliable. Specifically, the mean value depends on the range of values explored by the user. This lack of reliability is demonstrated in the unexpectedly negative correlation of the difference of mean values; while $H_1 - H_2$ also had unexpected negative correlations in Table I, its absolute value was lower than with $\mu_1 - \mu_2$.

V. DISCUSSION

This paper presents a novel method for assessing believability. We chose a top-down single-player shooter with an agent able to exhibit a different selection of behaviours, which allows us to test believability annotation in a complex environment without relying on human-like appearance or gestures from the artificial agent. Participants annotated their perceived moment-to-moment believability for two sessions and chose between the two agents, in terms of human-likeness in the given context. This allowed us to compare between continuous and discrete assessment techniques and between two different annotation tools.

In our results we see significant correlations between the participants' final choices and their annotations. We have explored different uncertainty thresholds with both tools and found BTrace to be closer to the ground truth. BTrace is showing higher correlations and more significant results than RankTrace in almost all thresholds. Interestingly, these findings match our previous study which sees both feature correlations and subsequent modelling to be more successful with BTrace as well [16]. In conclusion, this study encourages the use of

a discrete binary labelling protocol for character believability assessment. We believe BTrace is superior given how intuitive it is when asked whether the current context has 'human-like' behaviour or not. Its binary nature might be helping participants reduce the noise and uncertainty behind the term 'believable'. It is worth mentioning that the binary annotation of BTrace is closer to what we consider "ground truth" in this study as the binary believability preference between the two sessions. Combined with the different and more complex data processing steps taken to convert RankTrace annotations into high/low values, this may have impacted the results in favor of BTrace. Earlier work has shown that RankTrace is more suitable for measures such as arousal [13] where gradients and an increasing or decreasing affective state are more intuitive for a user. Since RankTrace operates on an unbounded value range, identifying high versus low values is more difficult for both the user and for the analysis method described in Section III-C. Perhaps a more relevant measure for RankTrace is not the number of high versus low instances but rather the instances where believability increases from one time window to the next versus the instances where it drops. Future work should explore the *gradient* of the trace [38] as an alternative measure of believability, testing it against the ground truth of binary preference.

With this paper we aim to put a stronger emphasis on character believability, given the higher need for realism in games [21]. This paper explored an additional method for its assessment, by introducing the use of an existing time continuous annotation tool previously used for reporting changes in arousal [13], [39], [41], [42]. Our results showcase the versatility of PAGAN as an annotation tool and indicate that the temporal dimension and context of believable AI agents assessment should not be ignored. This tool could be used in addition to the existing discrete methods, providing extra information and aiding in understanding context such as investigating time windows of high believability. The annotation protocols examined provide more techniques for human-computer interaction assessment.

However, this study comes with limitations. It is merely an introduction into moment-to-moment assessment of believability, with much to be explored before becoming an accepted framework. Data cleaning reduced significantly the number of participants, especially for RankTrace, and further data loss was observed with higher thresholds. In future work, more participants should be collected to further strengthen the results. Among the final participants, we also noticed a trend towards picking the second video in the boolean choice. By allowing players to test, play and annotate twice, several minutes pass between annotating the first video and the final questionnaire where this choice is made. Recency effects and limited memory capabilities may introduce noise to the data. Integrating the choice after both sessions first and annotating after should be investigated in future studies, in order to lower the time between the first video and the binary choice. Finally, this work explores one single-player game and future research will investigate the generality of this method by testing it with other games. These would require different levels of complexity, such as different genres, narrative or multi-player.

The methods of discrete and time-continuous believability assessment are not limited to games. Given the methods' versatility, this work can be adapted to virtual reality applications, verbal interactions, and many other aspects of agent assessment in human-computer interaction. A different interface, such as a button system in a controller, could also adapt this method for physical interactions (within robotics, for example). Moreover, other measures can be explored to further complement discrete methods, such as collecting bio-signals and digital imagery of the session.

Furthermore, despite being out of scope in this work, the telemetry collected can be used for agent modelling. It provides moment-to-moment context with matching believability annotations—effectively showing which agent's features are and are not believable to participants. The connection between this work and modelling opens possibilities into agent design and even agent AI. For instance, a human designer could adapt its agent's behaviour based on the perceived features that make it more believable. On the other hand, an AI process may aim to increase believability by using predictions as rewards in a reinforcement learning agent. Previous work on the use of pixels for arousal prediction [41] also show that the suggested collection of a session's screen image could allow for other techniques such as pixel-to-believability predictions.

VI. Conclusion

This paper introduces the idea of using two different time-continuous affect annotation tools for character believability assessment. In this novel methodology, we investigate how a discrete method—the commonly used subjective choice preference—can benefit from the addition of moment-to-moment agent assessment. Our data was provided by 89 participants in total, which includes two sessions of continuous annotations and a survey with a choice between both videos and demographic information. The results show a higher correlation between BTrace and the discrete method than RankTrace and are encouraging for the use of this new methodology in addition to the existing methods. We conclude that this study can establish an additional path for agent believability assessment. It also presents many options for future work, including experiments with more games, larger datasets, other types of human-computer interaction and modelling.

References

[1] C. Holmgård, M. C. Green, A. Liapis, and J. Togelius, "Automated playtesting with procedural personas through mcts with evolved heuristics," *IEEE Trans. on Games*, vol. 11, no. 4, pp. 352–362, 2018.

[2] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[3] C. Guerrero-Romero, S. M. Lucas, and D. Perez-Liebana, "Using a team of general AI algorithms to assist game design and testing," in *Proc. of the IEEE Conf. on Computational Intelligence and Games*. IEEE, 2018.

[4] C. L. Lisetti, "Believable agents, engagement, and health interventions," in *Proc. of the Intl. Conf. on Human-Computer Interaction*. Springer, 2011, pp. 425–432.

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

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

[7] H. Warpefelt, M. Johansson, and H. Verhagen, "Analyzing the believability of game character behavior using the game agent matrix," in *Proc. of the DiGRA Conf.*, 2013.

[8] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, p. 433, 1950.

[9] C. Even, A.-G. Bosser, and C. Buche, "Analysis of the protocols used to assess virtual players in multi-player computer games," in *Proc. of the Intl. Work-Conf. on Artificial Neural Networks*. Springer, 2017, pp. 657–668.

[10] C. Pacheco, L. Tokarchuk, and D. Pérez-Liébana, "Studying believability assessment in racing games," in *Proc. of the Intl. Conf. on the foundations of digital games*, 2018.

[11] E. Camilleri, G. N. Yannakakis, and A. Dingli, "Platformer level design for player believability," in *Proc. of the IEEE Conf. on Computational Intelligence and Games*, 2016.

[12] 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.

[13] D. Melhart, A. Liapis, and G. N. Yannakakis, "PAGAN: Video affect annotation made easy," in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*. IEEE, 2019, pp. 130–136.

[14] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2015, pp. 574–580.

[15] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation," in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2017, pp. 158–163.

[16] C. Pacheco, D. Melhart, A. Liapis, G. N. Yannakakis, and D. Perez-Liebana, "Trace it like you believe it: Time-continuous believability prediction," in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2021.

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

[18] A. B. Loyall, "Believable agents: Building interactive personalities." Carnegie-Mellon University, Tech. Rep., 1997.

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

[20] 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.

[21] S. Hamdy and D. King, "Affect and believability in game characters–a review of the use of affective computing in games," in *Proc. of the Annual Conf. on Simulation and AI in Computer Games*, 2017.

[22] P. Lankoski and S. Björk, "Gameplay design patterns for believable non-player characters." in *Proc. of the DiGRA Conf.*, 2007, pp. 416–423.

[23] P. Hingston, "A Turing test for computer game bots," *IEEE Trans. on Computational Intelligence and AI in Games*, vol. 1, no. 3, pp. 169–186, 2009.

[24] R. Aylett, M. Vala, P. Sequeira, and A. Paiva, "Fearnot!–an emergent narrative approach to virtual dramas for anti-bullying education," in *Proc. of the Intl. Conf. on Virtual Storytelling*. Springer, 2007, pp. 202–205.

[25] M. Ochs, N. Sabouret, and V. Corruble, "Simulation of the dynamics of nonplayer characters' emotions and social relations in games," *IEEE Trans. on Computational Intelligence and AI in Games*, vol. 1, no. 4, pp. 281–297, 2009.

[26] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[27] J. Gratch and S. Marsella, "Evaluating a computational model of emotion," *Autonomous Agents and Multi-Agent Systems*, vol. 11, no. 1, pp. 23–43, 2005.

[28] K. Bessiere, J. E. Newhagen, J. P. Robinson, and B. Shneiderman, "A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood," *Computers in human behavior*, vol. 22, no. 6, pp. 941–961, 2006.

[29] J. R. Searle, "Minds, Brains, and Programs," *Behavioral and brain sciences*, vol. 3, no. 3, pp. 417–424, 1980.

[30] P. Hayes and K. Ford, "Turing test considered harmful," in *Proc. of the Intl. joint conf. on Artificial intelligence*, 1995, pp. 972–977.

[31] D. Livingstone, "Turing's test and believable AI in games," *Computers in Entertainment (CIE)*, vol. 4, no. 1, pp. 6–19, 2006.

[32] R. Arrabales, A. Ledezma, and A. Sanchis, "Consscale: A pragmatic scale for measuring the level of consciousness in artificial agents," *Journal of Consciousness Studies*, vol. 17, no. 3-4, pp. 131–164, 2010.

[33] B. Mac Namee, "Proactive persistent agents-using situational intelligence to create support characters in character-centric computer games," Ph.D. dissertation, University of Dublin, Trinity College. Department of Computer Science, 2004.

[34] A. Bogdanovych, T. Trescak, and S. Simoff, "What makes virtual agents believable?" *Connection Science*, vol. 28, no. 1, pp. 83–108, 2016.

[35] J. D. Miles and R. Tashakkori, "Improving the believability of non-player characters in simulations," in *Proc. of the Conf. on Artificial General Intelligence*. Atlantis Press, 2009.

[36] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. of the automatic face and gesture recognition Conf.* IEEE, 2013.

[37] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, vol. 22, no. 140, p. 55, 1932.

[38] E. Camilleri, G. N. Yannakakis, and A. Liapis, "Towards general models of player affect," in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2017.

[39] D. Melhart, A. Liapis, and G. N. Yannakakis, "Towards general models of player experience: A study within genres," in *Proc. of the IEEE Conf. on Games*, 2021.

[40] K. Makantasis, A. Liapis, and G. N. Yannakakis, "The pixels and sounds of emotion: General-purpose representations of arousal in games," *IEEE Trans. on Affective Computing*, 2021.

[41] ——, "From pixels to affect: A study on games and player experience," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*, 2019.

[42] E. Xylakis, A. Liapis, and G. N. Yannakakis, "Architectural form and affect: A spatiotemporal study of arousal," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*, 2021.